



REC'D: 08 MAR 2004

WIPO

PCT

**Prioritätsbescheinigung über die Einreichung
einer Patentanmeldung**

Aktenzeichen:

102 60 805.9

Anmeldetag:

23. Dezember 2002

Anmelder/Inhaber:

GeneArt GmbH, 93053 Regensburg/DE

Bezeichnung:

Verfahren und Vorrichtung zum Optimieren einer
Nucleotidsequenz zur Expression eines Proteins

IPC:

C 07 H, C 12 P, C 12 N

**Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ur-
sprünglichen Unterlagen dieser Patentanmeldung.**

München, den 9. Februar 2004
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Deizeron

**PRIORITY
DOCUMENT**

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

BOEHMERT & BOEHMERT

ANWALTSSOZIELTÄT

Boehmert & Boehmert • P.O.B. 10 71 27 • D-28071 Bremen

Deutsches Patent- und Markenamt
Zweibrückenstraße 12

80297 München

DR.-ING. KARL BOEHMERT, PA (1898-1972)
DIPLO.-ING. ALBERT BOEHMERT, PA (1902-1997)
WILHELM J. H. STAHLBERG, RA, Bremen
DR.-ING. WALTER HOORMANN, PA*, Bremen
DIPLO.-PHYS. DR. HEINZ GÜDDAR, PA*, München
DR.-ING. ROLAND LIESSENG, PA*, München
WOLF-DIETER KUNTZE, RA, Bremen, Alicante
DIPLO.-PHYS. ROBERT MÜNZHUBER, PA (1912-1992)
DR. LUDWIG KOUKER, RA, Bremen
DR. (CHEM.) ANDREAS WÖHLER, PA*, Bremen
MICHAELA RUTH-DIERIG, RA, München
DIPLO.-PHYS. DR. MARION TONHARDT, PA*, Düsseldorf
DR. ANDREAS EBERT-WEIDENFELDER, RA, Bremen
DIPLO.-ING. EVA LIESSENG, PA*, München
DR. AGEL NORDEMANN, RA, Berlin
DIPLO.-PHYS. DR. DOROTHEE WEBER-BRULS, PA*, Frankfurt
DIPLO.-PHYS. DR. STEFAN SCHÖTZ, PA*, München
DR.-ING. MATTHIAS PHILIPP, PA*, Bielefeld
DR. MARTIN WIRTZ, RA, Düsseldorf
DR. DETMAR SCHÄFER, RA, Bremen
DR. JAN BIRND NORDEMANN, LL.M., RA, Berlin
DR. CHRISTIAN CZYCHOWSKI, RA, Berlin
DR. CARL-RICHARD HAARMANN, RA, München
DIPLO.-PHYS. CHRISTIAN W. APPELT, PA*, München

PA - Patentanwalt/Patent Attorney
RA - Rechtsanwalt/Attorney at Law
• - European Patent Attorney
• - Maître en Droit
• - Licencié en Droit
• - Diplôme d'Etudes Approfondies en Conception de Produits et Innovation
Als zugelassen zur Vertretung vor dem Europäischen Markenamt, Alicante
Professional Representation at the Community Trademark Office, Alicante

PROF. DR. WILHELM NORDEMANN, RA, Potsdam
DIPLO.-PHYS. EDUARD BAUMANN, PA*, Hohenkirchen
DR.-ING. GERALD KLOPPSCH, PA*, Düsseldorf
DIPLO.-ING. HANS W. GROENING, PA*, München
DIPLO.-ING. RIEGFRIED SCHIRMER, PA*, Bielefeld
DIPLO.-PHYS. LORENZ HANBINKEL, PA*, Potsdam
DIPLO.-ING. ANTON FRIEDRICH RIEDERER V. PAAR, PA*, Landshut
DIPLO.-ING. DR. JAN TONNER, PA, RA, Köln
DIPLO.-PHYS. CHRISTIAN BIEHL, PA*, Köln
DIPLO.-PHYS. DR.-ING. UWE MANASSE, PA*, Bremen
DIPLO.-PHYS. DR. THOMAS L. BITTNER, PA*, Berlin
DR. VOLKER SCHMIDT, M. Juris (Oxford), RA, München, Paris
DR. ANKE NORDEMANN-SCHÖTZ, RA*, Potsdam
DIPLO.-BIOL. DR. JAN B. KRAUSS, PA*, Berlin
DR. KLAUS TIM BROCKER, RA, Berlin
DR. ANDREAS DUSTMANN, LL.M., RA, Potsdam
DIPLO.-ING. NILS T. F. SCHMIDT, PA*, München, Paris
DR. FLORIAN SCHWAB, LL.M., RA*, München
DIPLO.-BIOCHEM. DR. MARKUS ENGELHARD, PA, München
DIPLO.-CHEM. DR. KARL-HEINZ B. MITTEN, PA*, Frankfurt
DIPLO.-ING. DR. STEFAN TARUTIS, PA, Düsseldorf
PASCAL DECKER, RA, Berlin
DIPLO.-CHEM. DR. VOLKER SCHOLZ, PA, Bremen
DIPLO.-CHEM. DR. JÖRK ZWICKER, PA, München
DR. CHRISTIAN MEISSNER, RA, München

In Zusammenarbeit mit/in cooperation with
DIPLO.-CHEM. DR. HANS ULRICH MAY, PA*, München

Ihr Zeichen
Your ref.

Ihr Schreiben
Your letter of

Unser Zeichen
Our ref.

Bremen,

Neuanmeldung

G30036

23. Dezember 2002

GeneArt GmbH
Josef-Engert-Straße 9
93053 Regensburg

Verfahren und Vorrichtung zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins

Ansprüche

1. Verfahren zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins, welches die folgenden auf einem Computer durchgeführten Schritte umfaßt:
 - Generieren einer ersten Testsequenz von n Codons, welche n aufeinanderfolgenden Aminosäuren in der Proteinsequenz entsprechen, wobei n eine natürli-

- 24.372 -

Hollerallee 32 • D-28209 Bremen • P.O.B. 10 71 27 • D-28071 Bremen • Telefon +49-421-34090 • Telefax +49-421-3491768

MÜNCHEN • BREMEN • BERLIN • DÜSSELDORF • FRANKFURT • BIELEFELD • POTSDAM • KIEL • PADERBORN • LANDSHUT • HOHENKIRCHEN • ALICANTE • PARIS

<http://www.boehmert.de>

e-mail: postmaster@boehmert.de

che Zahl und kleiner oder gleich N , der Zahl der Aminosäuren der Proteinsequenz, ist,

- Festlegen von m Optimierungspositionen in der Testsequenz, welche der Position von m Codons entsprechen, an denen die Besetzung mit einem Codon, bezogen auf die Testsequenz, optimiert werden soll, wobei $m \leq n$ und $m < N$ ist,
- Generieren einer oder mehrerer weiterer Testsequenzen aus der ersten Testsequenz, indem an einer oder mehreren der m Optimierungspositionen ein Codon der ersten Testsequenz durch ein anderes Codon ersetzt wird, welches dieselbe Aminosäure exprimiert,
- Bewerten jeder der Testsequenzen mit einer Gütefunktion und Ermitteln der hinsichtlich der Gütefunktion optimalen Testsequenz,
- Festlegen von p Codons der optimalen Testsequenz, welche sich an einer der m Optimierungspositionen befinden, als Ergebniscodons, welche die Codons der optimierten Nucleotidsequenz an den Positionen bilden, die der Position der besagten p Codons in der Testsequenz entspricht, wobei p eine natürliche Zahl und $p \leq m$ ist,
- Iterieren der vorangehenden Schritte, wobei in jedem Iterationsschritt die Testsequenz an den Positionen, welche Positionen von festgelegten Ergebniscodons in der optimierten Nucleotidsequenz entsprechen, das entsprechende Ergebniscodon enthält und die Optimierungspositionen von Positionen von Ergebniscodons verschieden sind.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß in einem oder mehreren Iterationsschritten die m Optimierungspositionen der Testsequenzen unmittelbar auf ein oder mehrere Ergebniscodons folgen, welche als Teil der optimierten Nucleotidsequenz festgelegt worden sind.
3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß in einem oder mehreren Iterationsschritten die p Codons, die als Ergebniscodons der optimierten Nucleotidsequenz festgelegt werden, p aufeinanderfolgende Codons sind.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß in einem Iterationsschritt Testsequenzen mit allen möglichen Codonbesetzungen für die m Optimierungspositionen aus der ersten Testsequenz generiert werden und die optimale Testsequenz unter diesen Testsequenzen ermittelt wird.
5. Verfahren nach einem der Ansprüche 1 bis 4, gekennzeichnet durch:
 - Bewerten jeder Testsequenz mit einer Gütefunktion,
 - Ermitteln eines Extremwertes innerhalb der Werte der Gütefunktion für alle in einem Iterationsschritt generierten Teilsequenzen,
 - Festlegen von p Codons der Testsequenz, welche dem extremalen Wert der Gewichtsfunktion entspricht, als Ergebniscodons an den entsprechenden Positionen, wobei p eine natürliche Zahl und $p \leq m$ ist.
6. Verfahren nach Anspruch 5, dadurch gekennzeichnet, daß die Gütefunktion eines oder mehrere der folgenden Kriterien berücksichtigt:
Codon usage für einen vorgegebenen Organismus, GC-Gehalt, repetitive Sequenzen, Sekundärstrukturen, invers komplementäre Sequenzwiederholungen und Sequenzmotive.
7. Verfahren nach Anspruch 6, dadurch gekennzeichnet, daß die Gütefunktion eine Funktion von verschiedenen Einzeltermen ist, die jeweils ein Kriterium aus der folgenden Liste von Kriterien bewerten:
Codon usage für einen vorgegebenen Organismus, GC-Gehalt, Sequenzmotive, repetitive Sequenzen, Sekundärstrukturen, invers komplementäre Sequenzwiederholungen.
8. Verfahren nach einem der Ansprüche 1 bis 6, dadurch gekennzeichnet, daß die Gütefunktion eines oder mehrere der folgenden Kriterien berücksichtigt:
 - Ausschluß von invers komplementären Sequenzidentitäten von mehr als 20 Nucleotiden zum Transkriptom eines vorgegebenen Organismus,

- Ausschluß von Homologiebereichen von mehr als 100 Basenpaaren zu einer vorgegebenen DNS-Sequenz,
 - Ausschluß von Homologiebereichen mit mehr als 90 % Ähnlichkeit der Nucleotidsequenz zu einer vorgegebenen DNS-Sequenz.
9. Verfahren nach einem der Ansprüche 1 bis 8, gekennzeichnet durch den Schritt des Synthetisierens der optimierten Nucleotidsequenz.
10. Verfahren nach Anspruch 9, dadurch gekennzeichnet, daß der Schritt des Synthetisierens der optimierten Nucleotidsequenz in einer Vorrichtung zum automatischen Synthetisieren von Nucleotidsequenzen stattfindet, welcher von dem Rechner angesteuert wird, der die Nucleotidsequenz optimiert.
11. Vorrichtung zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins, welche eine Recheneinrichtung aufweist, welche umfaßt:
- eine Einrichtung zum Generieren einer ersten Testsequenz von n Codons, welche n aufeinanderfolgenden Aminosäuren in der Proteinsequenz entsprechen, wobei n eine natürliche Zahl und kleiner oder gleich N , der Zahl der Aminosäuren der Proteinsequenz ist,
 - eine Einrichtung zum Festlegen von m Optimierungspositionen in der Testsequenz, welche der Position von m Codons entsprechen, an denen die Besetzung mit einem Codon, bezogen auf die Testsequenz, optimiert werden soll, wobei $m \leq n$ und $m < M$ ist,
 - eine Einrichtung zum Generieren einer oder mehrerer weiterer Testsequenzen aus der ersten Testsequenz, indem an einer oder mehreren der m Optimierungspositionen ein Codon der ersten Testsequenz durch ein anderes Codon ersetzt wird, welches dieselbe Aminosäure exprimiert,

- eine Einrichtung zum Bewerten jeder der Testsequenzen mit einer Gütefunktion und zum Ermitteln der hinsichtlich der Gütefunktion optimalen Testsequenz,
 - eine Einrichtung zum Festlegen von p Codons der optimalen Testsequenz, welche sich an einem der m Optimierungspositionen befinden, als Ergebniscodons, welche die Codons der optimierten Nucleotidsequenz an den Positionen bilden, die den Positionen der besagten p Codons in der Testsequenz entsprechen, wobei p eine natürliche Zahl und $p \leq m$ ist,
 - eine Einrichtung zum Iterieren der Schritte des Generierens mehrerer Testfunktionen, der Bewertung der Testsequenzen und des Festlegens von Ergebniscodons, wobei in jedem Iterationsschritt die Testsequenz an den Positionen, welche Positionen von festgelegten Ergebniscodons in der optimierten Nucleotidsequenz entsprechen, das entsprechende Ergebniscodon enthält und die Optimierungspositionen von Positionen von Ergebniscodons verschieden sind.
12. Vorrichtung nach Anspruch 11, gekennzeichnet durch eine Einrichtung zum Durchführen der Schritte eines Verfahrens nach einem der Ansprüche 1 bis 7.
 13. Vorrichtung nach einem der Ansprüche 11 oder 12, gekennzeichnet durch eine Vorrichtung zum automatischen Synthetisieren von Nucleotidsequenzen, welcher von dem Rechner so angesteuert wird, daß er die optimierte Nucleotidsequenz synthetisiert.
 14. Computerprogramm, welches von einem Computer ausführbaren Programmcode enthält, der, wenn er auf einem Computer ausgeführt wird, den Computer veranlaßt, ein Verfahren nach einem der Ansprüche 1 bis 8 durchzuführen.
 15. Computerprogramm nach Anspruch 14, wobei der Programmcode, wenn er auf einem Computer ausgeführt wird, eine Vorrichtung zum automatischen Synthetisieren von Nucleotidsequenzen veranlassen kann, die optimierte Nucleotidsequenz herzustellen.

16. Computerlesbarer Datenträger, auf welchem in computerlesbarer Form ein Programm nach einem der Ansprüche 14 oder 15 gespeichert ist.
17. Nukleinsäure, welche eine für ein Protein codierende Nucleotidsequenz umfaßt, erhältlich durch ein Verfahren nach Anspruch 9.
18. Nukleinsäure nach Anspruch 17, dadurch gekennzeichnet, daß diese eine Nucleotidsequenz umfaßt, welche in einem vorgegebenen Organismus für ein Protein codiert, wobei die besagte Nucleotidsequenz in dem natürlich vorkommenden Genom des Organismus nicht enthalten ist.
19. Nukleinsäure nach Anspruch 18, dadurch gekennzeichnet, daß der Organismus ausgewählt aus der folgenden Gruppe ist:
 - Viren, insbesondere Vaccinia-Viren,
 - Prokaryonten, insbesondere Escherichia coli, Caulobacter crescentus, Bacillus subtilis, Mycobacterium spec.,
 - Hefen, insbesondere Saccharomyces cerevisiae, Schizosaccharomyces pombe, Pichia pastoris, Pichia angusta,
 - Insekten, insbesondere Sprodoptera frugiperda, Drosophila spec,
 - Säuger, insbesondere Homo sapiens, Macaca mulata, Mus musculus, Bos taurus, Capra hircus, Ovis aries, Oryctolagus cuniculus, Rattus norvegicus, chinese hamster ovary,
 - Monokotyle Pflanzen, insbesondere Oryza sativa, Zea mays, Triticum aestivum
 - Dikotyle Pflanzen, insbesondere Glycin max, Gossypium hirsutum, Nicotiana tabacum, Arabidopsis thaliana, Solanum tuberosum.
20. Nukleinsäure nach einem der Ansprüche 1 bis 19, dadurch gekennzeichnet, daß das durch die Nucleotidsequenz codierte Protein eines der folgenden Proteine ist und/oder in einer der folgenden Proteinklassen fällt:

- Enzyme, insbesondere Polymerasen, Endonukleasen, Ligasen, Lipasen, Proteasen, Kinasen, Phosphatasen, Topoisomerasen,
 - Cytokine, Chemokine, Transkriptionsfaktoren, Oncogene,
 - Proteine aus thermophilen Organismen, aus cryophilen Organismen, aus halophilen Organismen, aus acidophilen Organismen, aus basophilen Organismen,
 - Proteine mit repetitiven Sequenzelementen, insbesondere strukturgebende Proteine,
 - Humane Antigene, insbesondere Tumorantigene, Tumormarker, Autoimmunantigene, diagnostische Marker,
 - Virale Antigene, insbesondere von HAV, HBV, HCV, HIV, SIV, FIV, HPV, Rinoviren, Influenzaviren, Herpesviren, Poliomaviren, Hendra Virus, Dengue Virus, AAV, Adenoviren, HTLV, RSV,
 - Antigene von parasitären Erregern, z.B. Protozoen, insbesondere Erreger von Malaria, Leishmania, Trypanosoma, Toxoplasmen, Amöba,
 - Antigene von bakteriellen Erregern oder Pathogene, insbesondere von den Genera Chlamydia, Staphylococci, Klebsiella, Streptococcus, Salmonella, Listeria, Borrelia, Escherichia coli,
 - Antigene von Organismen der Sicherheitstufe L4, insbesondere Bacillus anthracis, Ebola-Virus, Marburg-Virus, Pockenviren.
21. Nukleinsäure nach einem der Ansprüche 18 oder 19, dadurch gekennzeichnet, daß die Gütefunktion zumindest eines der folgenden Kriterien berücksichtigt:
- GC-Gehalt,
 - Codon Usage des vorgegebenen Organismus,
 - Ausschluß von invers komplementären Sequenzidentitäten von mehr als 20 Nucleotiden zum Transkriptom eines vorgegebenen Organismus,
 - vollständiger oder weitgehender Ausschluß von Homologiebereichen von mehr als 100 Basenpaaren zu einer vorgegebenen DNS-Sequenz,
 - vollständiger oder weitgehender Ausschluß von Homologiebereichen mit einer Ähnlichkeit von mehr als 90 % zu einer vorgegebenen DNS-Sequenz.

22. Vektor, umfassend eine Nukleinsäure nach einem der Ansprüche 17 bis 21.
23. Zelle, umfassend einen Vektor nach Anspruch 22 oder eine Nukleinsäure nach einem der Ansprüche 17 bis 21.
24. Organismus, umfassend mindestens eine Zelle nach Anspruch 23.

BOEHMERT & BOEHMERT

ANWALTSSOZIELTÄT

Boehmert & Boehmert • P.O.B. 10 71 27 • D-28071 Bremen

Deutsches Patent- und Markenamt
Zweibrückenstraße 12

80297 München

DR.-ING. KARL BOEHMERT, PA (1898-1977)
DIPLO.-ING. ALBERT BOEHMERT, PA (1902-1993)
WILHELM I. H. STAHLBERG, RA, Bremen
DR.-ING. WALTER HOORMANN, PA*, Bremen
DIPLO.-PHYS. DR. HEDZ. GODDAR, PA*, München
DR.-ING. ROLAND LIESEGANG, PA*, München
WOLF-DIETER KUNTZE, RA, Bremen, Altscheid
DIPLO.-PHYS. ROBERT MÜNZER, PA (1933-1992)
DR. LUDWIG KOUKER, RA, Bremen
DR. (CHEM.) ANDREAS WINKLER, PA*, Bremen
MICHAELA HUTH-DIERIG, RA, München
DIPLO.-PHYS. DR. MARION TONHARDT, PA*, Düsseldorf
DR. ANDREAS EBERT-WEIDENFELDER, RA, Bremen
DIPLO.-ING. EVA LIESEGANG, PA*, München
DR. AXEL NORDEMANN, RA, Berlin
DIPLO.-PHYS. DR. DOROTHEE WEBER-BRULS, PA*, Frankfurt
DIPLO.-PHYS. DR. STEFAN SCHÖPE, PA*, München
DR.-ING. MATTHIAS FLEPP, PA*, Bielefeld
DR. MARTIN WITZ, RA, Chemnitz
DR. DETMAR SCHÄFER, RA, Bremen
DR. JAN BERND NORDEMANN, LL.M., RA, Berlin
DR. CHRISTIAN CZYCHOWSKI, RA, Berlin
DR. CARL-RICHARD HAARMANN, RA, München
DIPLO.-PHYS. CHRISTIAN W. APFELT, PA*, München

PROF. DR. WILHELM NORDEMANN, RA, Potsdam
DIPLO.-PHYS. EDUARD BAUMANN, PA*, Hohenkirchen
DR.-ING. GERALD KLOPFCH, PA*, Chemnitz
DIPLO.-ING. HANS W. GROENING, PA*, München
DIPLO.-ING. KIEGFRIED SCHIRMER, PA*, Bielefeld
DIPLO.-PHYS. LORENZ HANSWINKEL, PA*, Potsdam
DIPLO.-ING. ANTON FRIEDRICH RIEDERER V. PAAR, PA*, Landshut
DIPLO.-ING. DR. JAN TONNER, PA, RA, Kiel
DIPLO.-PHYS. CHRISTIAN BIEHL, PA*, Kiel
DIPLO.-PHYS. DR.-ING. UWE MANASSE, PA*, Bremen
DIPLO.-PHYS. DR. THOMAS L. BITTNER, PA*, Berlin
DR. VOLKER SCHMIDT, M. Juris (Oxford), RA, München, Paris
DR. ANKE NORDEMANN-SCHIFFEL, RA*, Potsdam
DIPLO.-ING. DR. JAN B. KRAUSS, PA*, Berlin
DR. KLAUS TIM BRÖCKER, RA, Berlin
DR. ANDREAS DUSTMANN, LL.M., RA, Potsdam
DIPLO.-ING. NILS T. F. SCHMID, PA*, München, Paris
DR. FLORIAN SCHWAB, LL.M., RA*, München
DIPLO.-BIOCHEM. DR. MARKUS ENGELHARD, PA, München
DIPLO.-CHEM. DR. KARL-HEINZ B. METTEN, PA*, Frankfurt
DIPLO.-ING. DR. STEFAN TARUTIS, PA, Düsseldorf
PASCAL DECKER, RA, Berlin
DIPLO.-CHEM. DR. VOLKER SCHOLZ, PA, Bremen
DIPLO.-CHEM. DR. JÖRK ZWICKER, PA, München
DR. CHRISTIAN MEISSNER, RA, München

In Zusammenarbeit mit/in cooperation with
DIPLO.-CHEM. DR. HANS ULRICH MAY, PA*, München

PA = Patentanwalt/Patent Attorney
RA = Rechtsanwalt/Attorney at Law
• European Patent Attorney
• Maître en Droit
• Licencié en Droit
• Diplôme d'Etudes Approfondies en Conception de Produits et Innovation
Alle zugelassen zur Vertretung vor dem Europäischen Patentamt, Alicante
Professional Representation at the Community Trademark Office, Alicante

Ihr Zeichen
Your ref.

Ihr Schreiben
Your letter of

Unser Zeichen
Our ref.

Bremen,

Neuanmeldung

G30036

23. Dezember 2002

GeneArt GmbH
Josef-Engert-Straße 9
93053 Regensburg

Verfahren und Vorrichtung zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins

Die Erfindung betrifft allgemein die Erzeugung synthetischer DNS-Sequenzen und deren Verwendung zur Erzeugung von Proteinen, indem diese DNS-Sequenzen in ein Expressionssystem, zum Beispiel in einen Wirtsorganismus/eine Wirtszelle oder ein System für eine In-vitro-Expression eingebracht werden, der bzw. die das entsprechende Protein exprimiert. Sie betrifft insbesondere Verfahren, bei denen eine synthetische Nucleotidsequenz für das jeweilige Expressionssystem, also zum Beispiel für einen Organismus/für eine Wirtszelle, mit Hilfe eines Computers optimiert wird.

- 24.347 -

Hollerallee 32 • D-28209 Bremen • P.O.B. 10 71 27 • D-28071 Bremen • Telefon +49-421-34090 • Telefax +49-421-3491768

MÜNCHEN • BREMEN • BERLIN • DÜSSELDORF • FRANKFURT • BIELEFELD • POTSDAM • KIEL • PADERBORN • LANDSHUT • HOHENKIRCHEN • ALICANTE • PARIS

<http://www.boehmert.de>

e-mail: postmaster@boehmert.de

Eine Technik zur Herstellung und Synthetisierung von Proteinen ist das Klonen und Expressieren der dem Protein entsprechenden Gensequenz in heterologen Systemen, z.B. *Escherichia coli* oder Hefe. Natürlich vorkommende Gene sind für diesen Zweck allerdings häufig suboptimal. Da in einer DNS-Sequenz, die ein Protein exprimiert, jeweils ein Triplet von Basen (Codon) eine Aminosäure exprimiert, ist es möglich, eine künstliche DNS-Sequenz zur Expression des gewünschten Proteins zu synthetisieren und für das Klonen und Expressieren des Proteins zu verwenden. Ein Problem bei diesem Vorgehen besteht darin, daß einer vorgegebenen Aminosäuresequenz keine eindeutige Nucleotidsequenz entspricht. Dies wird als Degeneriertheit des genetischen Codes bezeichnet. Unterschiedliche Organismen verwenden Codons für die Expression einer Aminosäure mit unterschiedlicher Häufigkeit (sogenannte Codon usage). In der Regel gibt es in einem gegebenen Organismus ein Codon, das überwiegend verwendet wird und ein oder mehrere Codons, welche mit vergleichsweise geringer Häufigkeit von dem Organismus zur Expression der entsprechenden Aminosäure verwendet werden. Da die synthetisierte Nucleotidsequenz in einem bestimmten Organismus verwendet werden soll, sollte die Wahl der Codons an die Codon usage des entsprechenden Organismus angepaßt sein. Eine weitere wichtige Größe ist der GC-Gehalt (Gehalt der Basen Guanin und Cytosin in einer Sequenz). Weitere Faktoren, welche das Expressionsergebnis beeinflussen können, sind DNS-Motive und Wiederholungen oder invers komplementäre Wiederholungen in der Basensequenz. Bestimmte Basenabfolgen erzeugen in einem gegebenen Organismus bestimmte Funktionen, die innerhalb einer codierenden Sequenz nicht erwünscht sein können. Beispiele sind cis-aktive Sequenzmotive wie Spleißstellen oder Transkriptionsterminatoren. Das unbeabsichtigte Vorhandensein eines bestimmten Motivs kann die Expression reduzieren oder ganz unterdrücken oder sogar für den Wirtsorganismus eine toxische Wirkung haben. Sequenzwiederholungen können zu einer geringeren genetischen Stabilität führen und erschweren die Synthese repetitiver Abschnitte aufgrund der Gefahr von Fehlhybridisierungen. Invers komplementäre Wiederholungen können zur Bildung von unerwünschten Sekundärstrukturen auf der RNA-Ebene oder cruciformer Strukturen auf DNS-Ebene führen, welche die Transkription behindern und zu genetischer Instabilität führen, bzw. die Translationseffizienz negativ beeinflussen können.

Ein synthetisches Gen sollte daher hinsichtlich der Codon usage und des GC-Gehalts optimiert sein und andererseits die mit DNS-Motiven sowie Sequenzwiederholungen und invers komplementären Sequenzwiederholungen verbundenen Probleme weitgehend vermeiden. Diese Erfordernisse lassen sich in der Regel jedoch nicht gleichzeitig und in optimaler Weise erfüllen. Beispielsweise kann eine Optimierung auf die optimale Codon usage zu einer stark repetitiven Sequenz und einem erheblichen Abweichen von dem gewünschten GC-Gehalt führen. Es gilt daher, einen möglichst optimalen Kompromiß zwischen der Erfüllung der verschiedenen Erfordernisse herbeizuführen. Die große Anzahl von Aminosäuren in einem Protein führt jedoch zu einer kombinatorischen Explosion der Zahl der möglichen DNS-Sequenzen, welche – im Prinzip – das gewünschte Protein exprimieren können. Aus diesem Grund wurden verschiedene computergestützte Verfahren zum Ermitteln einer optimalen Codonsequenz vorgeschlagen.

P.S. Sarkar und Samir K. Brahmachari, *Nucleic Acids Research* 20 (1992), 5713 beschreiben Untersuchungen zur Rolle der Wahl der Codons bei der Bildung bestimmter räumlicher Strukturen einer DNS-Sequenz. Hierbei wurden alle möglichen degenerierten Nucleotidsequenzen generiert. Eine Bewertung der Sequenzen hinsichtlich des Vorhandenseins von strukturellen Motiven und strukturbildender Abschnitte erfolgte durch einen Computer unter Verwendung einer Wissensbasis. Die Verwendung einer Gütefunktion ist nicht offenbart.

D.M. Hoover und J. Lubkowski, *Nucleic Acid Research* 30 (2002), Nr. 10 e43 schlägt ein computergestütztes Verfahren vor, bei dem die Nucleotidsequenz in eine ungerade Anzahl von Abschnitten unterteilt wird, für die jeweils eine Gütefunktion (Score) berechnet wird. In die Gütefunktion gehen u.a. die Codon usage, die Möglichkeit der Bildung von Haarnadelstrukturen und die Abweichungen von der gewünschten Schmelztemperatur ein. Der Wert der Gütefunktion für die Gesamtsequenz bestimmt sich aus der Summe der Werte der Gütefunktion für die einzelnen Abschnitte. Die Besetzung mit Codons innerhalb eines Abschnittes wird durch ein sogenanntes Monte-Carlo-Verfahren optimiert. Dabei werden statistisch Codonpositionen ausgewählt, bei denen das Codon einer Ausgangssequenz durch ein statistisch aus-

gewähltes äquivalentes Codon ersetzt wird. Gleichzeitig werden in einer Iteration auch die Grenzen der Abschnitte neu definiert. Auf diese Weise wird eine vollständige Gensequenz statistisch generiert. Ist der Wert der Gütefunktion für die Gesamtsequenz kleiner als die bisherige Sequenz, wird die neue Sequenz beibehalten. Ist er größer, wird mit einer gewissen Wahrscheinlichkeit die neue Sequenz beibehalten, wobei diese Wahrscheinlichkeit durch eine Boltzmann-Statistik kontrolliert wird. Wenn sich innerhalb einer vorbestimmten Anzahl von Iterationen die Sequenz nicht ändert, wird diese Sequenz als optimale Sequenz bewertet.

Derartige statistische Verfahren haben den Nachteil, daß sie stark von der Wahl der Konvergenzkriterien abhängen.

Es ist die Aufgabe der Erfindung, ein alternatives Verfahren zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins zur Verfügung zu stellen, welches sich mit relativ geringem Speicherplatz und relativ geringer Rechenzeit auf einem Computer implementieren läßt und welches insbesondere Nachteile der statistischen Verfahren vermeidet.

Erfindungsgemäß wird diese Aufgabe durch ein Verfahren zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins gelöst, welches die folgenden auf einem Computer durchgeführten Schritte umfaßt:

- Generieren einer ersten Testsequenz von n Codons, welche n aufeinanderfolgenden Aminosäuren in der Proteinsequenz entsprechen, wobei n eine natürliche Zahl und kleiner oder gleich N , der Zahl der Aminosäuren der Proteinsequenz, ist,
- Festlegen von m Optimierungspositionen in der Testsequenz, welche der Position von m Codons, insbesondere von m aufeinanderfolgenden Codons, entsprechen, an denen die Besetzung mit einem Codon, bezogen auf die Testsequenz, optimiert werden soll, wobei $m \leq n$ und $m < N$ ist,
- Generieren einer oder mehrerer weiterer Testsequenzen aus der ersten Testsequenz, indem an einer oder mehreren der m Optimierungspositionen ein Codon der ersten

Testsequenz durch ein anderes Codon ersetzt wird, welches dieselbe Aminosäure exprimiert,

- Bewerten jeder der Testsequenzen mit einer Gütefunktion und Ermitteln der hinsichtlich der Gütefunktion optimalen Testsequenz,
- Festlegen von p Codons der optimalen Testsequenz, welche sich an einer der m Optimierungspositionen befinden, als Ergebniscodons, welche die Codons der optimierten Nucleotidsequenz an den Positionen bilden, die der Position der besagten p Codons in der Testsequenz entspricht, wobei p eine natürliche Zahl und $p \leq m$ ist,
- Iterieren der vorangehenden Schritte, wobei in jedem Iterationsschritt die Testsequenz an den Positionen, welche Positionen von festgelegten Ergebniscodons in der optimierten Nucleotidsequenz entsprechen, das entsprechende Ergebniscodon enthält und die Optimierungspositionen von Positionen von Ergebniscodons verschieden sind.

Gemäß der bevorzugten Ausführungsform der Erfindung werden die vorangehend genannten Schritte so oft iteriert, bis alle Codons der optimierten Nucleotidsequenz festgelegt, d.h. mit Ergebniscodons besetzt worden sind.

Erfindungsgemäß wird also die Sequenz nicht insgesamt, sondern sukzessiv auf Teilbereichen optimiert. Die in einem Iterationsschritt als optimal festgelegten p Ergebniscodons werden in den nachfolgenden Iterationsschritten nicht mehr verändert und vielmehr bei den jeweiligen Optimierungsschritten als gegeben vorausgesetzt. Vorzugsweise ist die Anzahl der Ergebniscodons, welche auf diese Weise für die weiteren Iterationen festgelegt und als vorgegeben behandelt werden, kleiner als die Anzahl m der Optimierungspositionen, an denen in einem Iterationsschritt die Codons variiert werden. Zumindest in der Mehrzahl der Iterationsschritte, bei einer besonderen Ausführungsform bei allen Iterationsschritten außer dem ersten, ist wiederum m kleiner als die Zahl der Codons der Testsequenz (n). Dies gestattet es, nicht nur lokale Effekte auf den m variierten Positionen, sondern auch längerreichweitige Korrelationen, z.B. im Zusammenhang mit der Entstehung von RNA-Sekundärstrukturen, zu berücksichtigen.

Gemäß den derzeit bevorzugten Ausführungsformen liegt m im Bereich von 3 bis 20, vorzugsweise im Bereich von 5 bis 10. Bei dieser Wahl dieses Parameters kann die Variation der Codons mit einem akzeptablen Aufwand an Speicher und Rechenzeit durchgeführt werden und gleichzeitig eine gute Optimierung der Sequenz erreicht werden.

Gemäß einer Ausführungsform muß m in den verschiedenen Iterationsschritten nicht gleich sein, sondern kann vielmehr auch in unterschiedlichen Iterationsschritten verschieden sein. Es kann auch vorgesehen sein, in einem Iterationsschritt die Variation der Testsequenz für verschiedene Werte von m durchzuführen und ggf. nur das Optimierungsergebnis für einen Wert von m zu berücksichtigen, um Einflüsse der Größe m auf das Optimierungsergebnis zu reduzieren bzw. um zu überprüfen, ob eine Vergrößerung der Zahl m zu einer Änderung des Ergebnisses führt.

Gemäß der bevorzugten Ausführungsform sind die m Optimierungspositionen oder zumindest ein Teil davon zusammenhängend und bilden somit ein Variationsfenster in der Testsequenz, auf welchem die Codonbesetzung variiert wird.

Die Erfindung kann insbesondere vorsehen, daß in zwei oder mehr aufeinanderfolgenden Iterationsschritten ein Teil der m Optimierungspositionen, auf welchen die Codons variiert werden, identisch sind. Sind die m Positionen zusammenhängend, bedeutet dies, daß das Variationsfenster bei einem Iterationsschritt mit dem Variationsfenster eines vorangehenden Iterationsschrittes überlappt.

Die Erfindung kann vorsehen, daß in einem oder mehreren Iterationsschritten die m Optimierungspositionen der Testsequenzen unmittelbar auf ein oder mehrere Ergebniscodons folgen, welche als Teil der optimierten Nucleotidsequenz festgelegt worden sind.

Die Erfindung kann ebenfalls vorsehen, daß in einem oder mehreren Iterationsschritten die p Codons, die als Ergebniscodons der optimierten Nucleotidsequenz festgelegt werden, p aufeinanderfolgende Codons sind, die vorzugsweise unmittelbar auf ein oder mehrere Ergeb-

niscodons folgen, welche als Teil der optimierten Nucleotidsequenz in einem früheren Schritt festgelegt worden sind.

Die Erfindung kann vorsehen, daß die Nucleotidsequenz von einem ihrer Enden her optimiert wird. Insbesondere kann die Erfindung vorsehen, daß in jedem Iterationsschritt die Länge der Testsequenz des vorherigen Iterationsschritts um eine bestimmte Anzahl Codons, die in unterschiedlichen Iterationen verschieden sein kann, vergrößert wird, bis $n = N$ ist. Ist $n = N$ und die Zahl derjenigen Positionen, die in der Testsequenz nicht mit Ergebniscodons besetzt sind, kleiner oder gleich dem Wert von m , der in den vorangehenden Iterationen verwendet wurde, oder liegt diese Zahl, bei Verwendung unterschiedlicher Werte von m in verschiedenen Iterationen, im Bereich der in Frage kommenden Werte von m , kann in dem entsprechenden Iterationsschritt $p = m$ gesetzt werden, wobei m gleichzeitig die Zahl der noch nicht festgelegten Codons ist. Die als optimal aufgefundene Besetzung der Optimierungspositionen wird dann für die Ergebniscodons an diesen Optimierungspositionen übernommen. Dies gilt insbesondere dann, wenn für jede mögliche Kombination von Besetzungen der Optimierungspositionen eine Testsequenz generiert wird.

Es kann jedoch auch vorgesehen sein, daß der Bereich der Testsequenz innerhalb der gesamten Sequenz in einem Iterationsschritt nicht oder nicht vollständig den Bereich einer Testsequenz in einem vorherigen Iterationsschritt umfaßt. Beispielsweise kann die Testsequenz selbst ein Fenster auf der Gesamtsequenz, z.B. ein Fenster fester Länge, bilden, das im Laufe der verschiedenen Iterationen auf der Gesamtsequenz verschoben wird.

Gemäß einer bevorzugten Ausführungsform wird die Testsequenz nach jedem Schritt um p Codons verlängert, wobei insbesondere m für alle Iterationsschritte konstant sein kann.

Analog zu der vorangehend beschriebenen Ausführungsform der Erfindung kann auch vorgesehen sein, daß die Nucleotidsequenz von einer Stelle in ihrem Inneren her optimiert wird. Dies kann z.B. in der Art geschehen, daß eine anfängliche Testsequenz, welche einem Bereich im Inneren der zu optimierenden Nucleotidsequenz entspricht, zunächst nach einer Seite suk-

zessiv vergrößert wird, bis das Ende der zu optimierenden Nucleotidsequenz oder ein anderer vorgegebener Punkt der zu optimierenden Nucleotidsequenz erreicht ist, und dann die Testsequenz zu der anderen Seite hin vergrößert wird, bis dort das andere Ende der zu optimierenden Nucleotidsequenz oder ein anderer vorgegebener Punkt der zu optimierenden Nucleotidsequenz erreicht ist.

Die Erfindung kann auch vorsehen, daß die Testsequenzen in einem Iterationsschritt aus einer optimierten oder anderweitig festgelegten Teilsequenz der Länge q und zwei auf beiden Seiten daran anschließenden Variationsbereichen mit einer Länge von m_1 bzw. m_2 Codons besteht, wobei $q+m_1+m_2 = n$ gilt. Die Besetzung der Variationsbereiche kann für beide Variationsbereiche gemeinsam optimiert werden, indem die Codons auf den m_1 und m_2 Plätzen gleichzeitig variiert und optimiert werden. Vorzugsweise werden in einem solchen Fall in jedem Iterationsschritt p_1 und p_2 Codons in dem ersten und zweiten Variationsbereich festgelegt, welche der weiteren Iteration als gegeben zugrunde gelegt werden. Es kann jedoch auch vorgesehen sein, daß die beiden Variationsbereiche unabhängig voneinander variiert und optimiert werden. Beispielsweise kann vorgesehen sein, daß die Besetzung nur in einem der beiden Variationsbereiche variiert wird und nur in dem einen Bereich Codons festgelegt werden, bevor die Variation und Optimierung in den zweiten Bereich stattfindet. In diesem Fall werden die p_1 festgelegten Codons in dem ersten Bereich bei der Optimierung des zweiten Bereichs als gegeben vorausgesetzt. Dieses Vorgehen ist dann sinnvoll, wenn allenfalls geringe Korrelationen zwischen den beiden Bereichen zu erwarten sind.

Gemäß dieser Ausführungsform kann vorgesehen sein, daß die Nucleotidsequenz von einem Punkt oder einem Bereich im Inneren der Sequenz ausgehend optimiert wird.

Die Erfindung kann insbesondere vorsehen, daß in jedem Iterationsschritt der Bereich der Testsequenz auf der Gesamtsequenz den Bereich der Testsequenzen in allen vorangehenden Iterationsschritten umfaßt und der Bereich einer Testsequenz in zumindest einigen der vorangehenden Iterationsschritte jeweils im Inneren oder jeweils am Rand des Bereichs der Testsequenz in dem aktuellen Iterationsschritt liegt.

Die Erfindung kann vorsehen, daß die Nucleotidsequenz auf verschiedenen Teilbereichen unabhängig optimiert wird. Die optimierte Nucleotidsequenz kann dann die Kombination der verschiedenen optimierten Teilsequenzen sein. Es kann auch vorgesehen sein, daß zumindest ein Teil der jeweiligen Ergebniscodons von zwei oder mehr optimierten Teilbereichen als Bestandteil einer Testsequenz in einer oder mehreren Iterationen verwendet wird.

Gemäß einer bevorzugten Ausführungsform der Erfindung ist vorgesehen, daß in einem Iterationsschritt Testsequenzen mit allen möglichen Codonbesetzungen für die m Optimierungspositionen aus der ersten Testsequenz generiert werden und die optimale Testsequenz unter allen möglichen Testsequenzen, bei denen an einer oder mehreren der m Optimierungspositionen ein Codon durch ein anderes Codon, welches dieselbe Aminosäure exprimiert, ersetzt wurde, ermittelt wird.

Gemäß einer Ausführungsform der Erfindung ist die zum Bewerten der Testsequenzen verwendete Gütefunktion bei allen oder zumindest der Mehrzahl der Iterationen gleich. Die Erfindung kann jedoch auch vorsehen, unterschiedliche Gütefunktionen in unterschiedlichen Iterationen, zum Beispiel in Abhängigkeit von der Länge der Testsequenzen, zu verwenden.

Das erfindungsgemäße Verfahren kann insbesondere die folgenden Schritte umfassen:

- Bewerten jeder Testsequenz mit einer Gütefunktion,
- Ermitteln eines Extremwertes innerhalb der Werte der Gütefunktion für alle in einem Iterationsschritt generierten Teilsequenzen,
- Festlegen von p Codons der Testsequenz, welche dem extremalen Wert der Gütefunktion entspricht, als Ergebniscodons an den entsprechenden Positionen, wobei p eine natürliche Zahl und $p \leq m$ ist.

Die Gütefunktion kann so definiert sein, daß die Sequenz entweder umso näher an dem Optimum liegt, je größer der Wert der Gütefunktion ist, oder umso näher an dem Optimum liegt, je kleiner ihr Wert ist. Entsprechend wird man bei dem Schritt des Ermitteln des Extrem-

wertes das Minimum oder das Maximum der Gütefunktion unter den generierten Codonsequenzen ermitteln.

Die Erfindung kann vorsehen, daß die Gütefunktion eines oder mehrere der folgenden Kriterien berücksichtigt:

Codon usage für einen vorgegebenen Organismus, GC-Gehalt, Sequenzmotive, repetitive Sequenzen, Sekundärstrukturen, inverse Repeats.

Die Erfindung kann insbesondere vorsehen, daß die Gütefunktion eines oder mehrere der folgenden Kriterien berücksichtigt:

- cis-aktive Sequenz-Motive, insbesondere DNS/Protein-Interaktionsbindestellen und RNS/Protein-Interaktionsbindestellen, bevorzugt Spleißmotive, Transkriptionsfaktorbindestellen, Transkriptionsterminatorenbindestellen, Polyadenylierungssignale, Endonucleaseerkennungssequenzen, immunomodulatorische DNS-Motive, Ribosomenbindestellen, Erkennungssequenzen für rekombinationsaktive Enzyme, Erkennungssequenzen für DNS-modifizierende Enzyme, Erkennungssequenzen für RNS-modifizierende Enzyme, Sequenzmotive, die in einem vorgegebenen Organismus unterrepräsentiert sind.

Die Erfindung kann auch vorsehen, daß die Gütefunktion eines oder mehrere der folgenden Kriterien berücksichtigt:

- Ausschluß oder weitgehender Ausschluß von invers komplementären Sequenzidentitäten von mehr als 20 Nukleotiden zum Transkriptom eines vorgegebenen Organismus,
- Ausschluß oder weitgehender Ausschluß von Homologiebereichen von mehr als 1.000 Basenpaaren, bevorzugt 500 Basenpaaren, stärker bevorzugt 100 Basenpaaren zu einer vorgegebenen DNS-Sequenz, zum Beispiel zu dem Genom eines vorgegebenen Organismus oder zu der DNS-Sequenz eines vorgegebenen Vektorkonstrukts.

Das erste dieser beiden Kriterien betrifft den Ausschluß des als RNA-Indifferenz bekannten Mechanismus, mit dem ein Organismus RNA-Sequenzen mit mehr als 20 Nukleotiden exakter Identität zu einer anderen RNA-Sequenz eliminiert oder deaktiviert. Mit dem zweiten Kriterium soll verhindert werden, daß eine Rekombination, das heißt ein Einbau der Sequenz in das Erbgut des Organismus, oder eine Mobilisierung von DNS-Sequenzen durch Rekombination mit anderen Vektoren stattfindet. Beide Kriterien können als absolute Ausschlußkriterien verwendet werden, d.h. Sequenzen, bei denen eines oder beide dieser Kriterien erfüllt sind, werden nicht berücksichtigt. Die Erfindung kann auch, wie nachfolgend noch genauer im Zusammenhang mit Sequenzmotiven erläutert wird, vorsehen, daß diesen Kriterien ein Gewicht zugeordnet ist, das betragsmäßig größer ist als der größte Beitrag von Kriterien zu der Gütefunktion, welche keine Ausschlußkriterien sind.

Die Erfindung kann auch, gegebenenfalls zusammen mit anderen Kriterien, das Kriterium vorsehen, daß keine Homologiebereiche erzeugt werden, die mehr als 90 % Ähnlichkeit und/oder 99 % Identität zu einer vorgegebenen DNS-Sequenz, zum Beispiel zu der entsprechenden Genomsequenz des vorgegebenen Organismus oder zu der DNS-Sequenz eines vorgegebenen Vektorkonstrukts aufweisen. Auch dieses Kriterium kann entweder als absolutes Ausschlußkriterium realisiert sein oder in einer Weise, daß es einen sehr großen Beitrag zu der Gütefunktion leistet, welcher den Beitrag anderer Kriterien, die nicht Ausschlußkriterien sind, überwiegt.

Insbesondere kann vorgesehen sein, daß die Gütefunktion eine Funktion von verschiedenen Einzeltermen, insbesondere eine Summe von Einzeltermen ist, die jeweils ein Kriterium aus der folgenden Liste von Kriterien bewerten:

Codon usage für einen vorgegebenen Organismus, GC-Gehalt, DNS – Motive, repetitive Sequenzen, Sekundärstrukturen, inverse Repeats.

Die besagte Funktion von Einzeltermen kann insbesondere eine Linearkombination von Einzeltermen oder eine rationale Funktion von Einzeltermen sein. Die genannten Kriterien müs-

sen nicht notwendigerweise vollständig in der Gewichtsfunktion berücksichtigt werden. Es kann auch nur ein Teil der Kriterien in der Gewichtsfunktion verwendet werden.

Die verschiedenen Einzelterme in der besagten Funktion werden nachfolgend Kriteriumsgewichte genannt.

Die Erfindung kann vorsehen, daß das Kriteriumsgewicht betreffend die Codon Usage (CU Score) proportional zu $\sum_i f_{ci}/f_{cmaxi}$ ist, wobei

- f_{ci} die Häufigkeit des an der Stelle i der Testsequenz gesetzten Codons für den betreffenden Organismus zur Expression der Aminosäure an der Stelle i der Aminosäuresequenz des zu exprimierenden Proteins ist und
- f_{cmaxi} die Häufigkeit des Codons ist, welches in dem entsprechenden Organismus am häufigsten die Aminosäure an der Stelle i exprimiert.

Das Maß f_{ci}/f_{cmaxi} ist als „Relative Adaptiveness“ bekannt (vgl. P. M. Sharp, W. H. Li, Nucleic Acids Research 15 (3) (1987), 1281 bis 1295).

Das lokale Gewicht des am häufigsten vorkommenden Codons wird dabei, unabhängig von der absoluten Häufigkeit, mit der dieses Codon vorkommt, auf einen bestimmten Wert, zum Beispiel 1, gesetzt. Damit wird vermieden, daß die Positionen, an denen nur wenige Codons zur Auswahl stehen, stärker zu dem Gesamtgewicht beitragen als diejenigen, an denen eine größere Anzahl von Codons zur Expression der Aminosäure zur Auswahl stehen. Der Index i kann über die gesamten n Codons der Testsequenz oder einen Teil davon laufen. Insbesondere kann in einer Ausführungsform vorgesehen sein, daß i nur über die m Codons der Optimierungspositionen läuft.

Die Erfindung kann vorsehen, daß das Kriteriumsgewicht betreffend die Codonusage nur für die m Ordnungspositionen verwendet wird.

Anstelle der Relative Adaptiveness kann auch die sogenannte RSCU (Relative Synonymous Codon Usage; vgl. P. M. Sharp, W. H. Li, a.a.O.) verwendet werden. Die RSCU für eine Codonposition ist definiert durch

$$RSCU_{ci} = f_{ci}d_i / (\sum_c f_{ci})$$

definiert, wobei die Summe im Nenner über alle Codons läuft, welche die Aminosäure an der Stelle i exprimieren und wobei d_i die Zahl der Codons angibt, welche die besagte Aminosäure exprimieren. Um ein Kriteriumsgewicht auf der Grundlage der RSCU zu definieren, kann vorgesehen sein, daß die RSCU für die jeweilige Testsequenz über alle Codons der Testsequenz oder einen Teil davon, insbesondere über die m -Codons der Optimierungspositionen, summiert wird. Der Unterschied zu dem von der Relative Adaptiveness abgeleiteten Kriteriumsgewicht besteht darin, daß bei dieser Gewichtung jede Codonposition mit dem Grad der Degeneriertheit, d_i , gewichtet wird, so daß solche Positionen, an denen mehr Codons zur Auswahl stehen, stärker in das Kriteriumsgewicht eingehen als solche Positionen, an denen nur wenige Codons oder sogar nur ein einziges Codon zur Auswahl stehen.

Bei den vorangehend beschriebenen Kriteriumsgewichten für die Codon-Usage wurde das arithmetische Mittel über die lokalen Gewichte (Relative Adaptiveness, RSCU) gebildet.

Es kann auch vorgesehen sein, daß das Kriteriumsgewicht betreffend die Codon-Usage proportional zu den geometrischen Mittel der lokalen Relative Adaptiveness bzw. der lokalen RSCU ist, so daß also gilt

$$CUScore = K(\prod_i RSCU_i)^{1/L}$$

oder

$$CUScore = K (\prod_i f_{ci}/f_{cmaxi})^{1/L}$$

ist, wobei K ein Skalierungsfaktor ist und L die Anzahl der Positionen ist, über welche das Produkt gebildet wird. Auch hier kann das Produkt wieder über die gesamte Testsequenz oder einen Teil, insbesondere über die m Optimierungspositionen, gebildet werden.

In diesem Zusammenhang stellt die Erfindung auch ein Verfahren zum Optimieren einer Nukleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins zur Verfügung, welches die folgenden auf einem Computer durchgeführten Schritte umfaßt:

- Generieren einer oder mehrerer Testsequenzen von n Codons, welche n aufeinanderfolgende Aminosäuren in der Proteinsequenz entsprechen, wobei n eine natürlich Zahl kleiner oder gleich N, der Zahl der Aminosäuren der Proteinsequenz, ist,
- Bewerten der einen oder mehreren Testsequenzen auf der Grundlage einer Gütefunktion, welche ein geometrisches oder arithmetisches Mittel der Relative Adaptiveness oder der RSCU über eine Anzahl von L Codonpositionen enthält, wobei L kleiner oder gleich N ist,
- Generierung einer oder mehrerer neuer Testsequenzen in Abhängigkeit von dem Ergebnis der besagten Bewertung.

Dabei kann die Generierung einer oder mehrerer neuer Testfunktionen in der oben beschriebenen Weise derart erfolgen, daß die neuen Testsequenzen eine bestimmte Anzahl aufgrund der vorangehenden Iterationen festgelegte Ergebniscodons enthalten, aber z.B. auch so, daß eine bestimmte Testsequenz mit einer bestimmten Wahrscheinlichkeit, die von dem Wert der Gütefunktion abhängt, als Grundlage für weitere Iterationen, insbesondere die weitere Erzeugung von Testsequenzen, verwendet wird, wie dies bei Monte-Carlo-Verfahren der Fall ist.

Während die Qualität eines Codons bei den obengenannten Verfahren durch die Nutzungshäufigkeit im Transkriptom oder einem Gen-Referenzset des Expressionsorganismus definiert wird, kann die Güte eines bestimmten Codons alternativ auch durch die biophysikalischen Eigenschaften des Codons selbst beschrieben werden. So ist zum Beispiel bekannt, daß Codons mit einer mittleren Codon-Anticodon-Bindungsenergie besonders effizient translatiert

werden. Als Maß für die translatorische Effizienz einer Testsequenz kann daher zum Beispiel der P2-Index verwendet werden, welcher das Verhältnis der Häufigkeit von Codons mit mittlerer Bindungsenergie und Codons mit extrem starker bzw. schwacher Bindungsenergie angibt. Alternativ können auch experimentell oder durch theoretische Berechnungen gewonnene Daten zur translatorischen Effizienz oder translationsgenauigkeit eines Codons zur Gütebewertung genutzt werden. Die oben genannten Bewertungskriterien können besonders dann von Vorteil sein, wenn die tRNA-Frequenzen des Expressionssystems nicht berücksichtigt werden müssen, da diese wie zum Beispiel bei in Vitro-Translationssystemen vom Experimentator festgelegt werden können.

Die Erfindung kann vorsehen, daß das Kriteriumsgewicht betreffend den GC-Gehalt (GCScore) eine Funktion des Betrags der Differenz des ermittelten GC-Gehalts der Teilsequenz, GCG, zu dem optimalen GC-Gehalt, GCG_{opt} ist, wobei unter dem GG-Gehalt der relative Anteil von Guanin und Cytosin, zum Beispiel in Form eines bestimmten prozentualen Anteils, zu verstehen ist.

Insbesondere kann das Kriteriumsgewicht GCScore die folgende Form haben:

$$GCScore = |\overline{GCG} - GCG_{opt}|^g \cdot h$$

wobei

\overline{GCG} der tatsächliche GC - Gehalt der Testsequenz oder eines vorbestimmten Teils der Testsequenz, GCG, oder der mittlere GC - Gehalt der Testsequenz oder eines vorbestimmten Teils der Testsequenz, $\langle GCG \rangle$, ist,

GCG_{opt} der gewünschte (optimale) GC - Gehalt ist,

g eine positive reelle Zahl, vorzugsweise im Bereich von 1 bis 3 , insbesondere 1,3 ist,

h eine positive reelle Zahl ist.

Der Faktor h ist im wesentlichen ein Gewichtungsfaktor, welcher das relative Gewicht des Kriteriumsgewichts GC-Score gegenüber den anderen Kriteriumsgewichten definiert. Vorzugsweise wird h so gewählt, daß der Betrag des maximal erreichbaren Wertes von GC-Score in einem Bereich von einem Hundertstel bis zu dem Hundertfachen eines anderen Kriteriumsgewichtes, insbesondere aller Kriteriumsgewichte, welche keine Ausschlußbedingung darstellen, wie zum Beispiel die Gewichte für ein erwünschtes bzw. unerwünschtes Sequenzmotiv, beträgt.

Zur Bestimmung des mittleren GC-Gehalts kann vorgesehen sein, daß ein auf eine bestimmte Basenposition bezogener lokaler GC-Gehalt durch den GC-Gehalt auf einem Fenster bestimmter Größe definiert wird, welches diese Base enthält und welches insbesondere bezüglich dieser Base zentriert sein kann. Dieser lokale GC-Gehalt wird dann über die Testsequenz oder einen Teilbereich der Testsequenz, insbesondere über die m Optimierungspositionen, gemittelt, wobei auch hier sowohl ein arithmetisches als auch ein geometrisches Mittel verwendet werden kann. Verwendet man einen auf diese Weise definierten mittleren GC-Gehalt, ergeben sich geringere Schwankungen zwischen Testsequenzen mit einer verschiedenen Länge n .

Die Erfindung kann vorsehen, daß der GC-Gehalt über einem Fenster ermittelt wird, welches größer als der Bereich der m Optimierungspositionen ist und diesen einschließt. Wenn die Optimierungspositionen ein zusammenhängendes Variationsfenster bilden, kann vorgesehen sein, daß b Basen vor und/oder nach dem Variationsfenster in die Bestimmung des Kriteriumsgewichts für den GC-Gehalt (GC-Score) einbezogen werden, wobei b in einem Bereich von 15 bis 45 Basen (entspricht 5 bis 15 Codons), vorzugsweise in einem Bereich von 20 bis 30 Basen liegen kann.

Die Erfindung kann weiterhin vorsehen, daß, soweit die Gütefunktion maximiert wird, bei der Ermittlung des Werts der Gütefunktion für jedes Vorkommen eines nicht erlaubten oder unerwünschten Sequenzmotivs ein fester Betrag abgezogen und für jedes erwünschte oder geforderte Motiv ein fester Betrag addiert wird (bei einer Minimierung der Gütefunktion verhält

es sich umgekehrt). Bei unerwünschten oder geforderten Motiven kann dieser Betrag deutlich größer sein als alle anderen Kriteriumsgewichte, so daß die anderen Kriterien demgegenüber nicht ins Gewicht fallen. Dadurch wird ein Ausschlußkriterium realisiert, während gleichzeitig eine Differenzierung danach stattfindet, ob ein Motiv einmal oder mehrfach aufgetreten ist. Ebenso läßt sich jedoch auch dann noch eine sinnvolle Gütefunktion definieren bzw. eine Bewertung der Testsequenzen mit der Gütefunktion durchführen, wenn die Bedingung hinsichtlich des Sequenzmotivs (Nichtvorhandensein eines bestimmten Motivs/Vorhandensein eines bestimmten Motivs) für alle in einem Iterationsschritt erzeugten Testsequenzen nicht erfüllt werden kann. Dies wird insbesondere dann der Fall sein, wenn die Länge n der Testsequenzen relativ klein gegenüber N ist, da aufgrund der vorgegebenen Aminosäuren der Proteinsequenz ein bestimmtes Motiv häufig erst bei größeren n auftreten kann.

Die Erfindung kann weiterhin vorsehen, daß die gesamte Testsequenz oder ein Teil davon daraufhin überprüft wird, ob bestimmte partielle Sequenzabschnitte oder zu bestimmten partiellen Sequenzabschnitten ähnliche Sequenzabschnitte in einem anderen Bereich der Testsequenz oder eines gegebenen Bereichs der Testsequenz auftreten oder ob bestimmte partielle Sequenzabschnitte oder zu bestimmten partiellen Sequenzabschnitten ähnliche Sequenzabschnitte in der invers komplementären Testsequenz oder eines Teils der invers komplementären Testsequenz vorkommen, und in Abhängigkeit hiervon ein Kriteriumsgewicht für Sequenzwiederholungen (repeats) und/oder inverse Sequenzwiederholungen (inverse repeats) berechnet wird. Im Regelfall wird dabei die Sequenz nicht nur darauf überprüft, ob ein bestimmter Sequenzabschnitt identisch in der Testsequenz bzw. der invers-komplementären Testsequenz bzw. eines Teilbereichs davon enthalten ist, sondern auch darauf, ob eine ähnliche, also nur teilweise übereinstimmende Sequenz in der Testsequenz bzw. der invers-komplementären Testsequenz bzw. eines Teils davon enthalten ist. Algorithmen zum Auffinden von globalen Übereinstimmungen (Global-Alignment-Algorithmen) oder lokalen Übereinstimmungen (Local Alignment-Algorithmen) zweier Sequenzen sind in der Bioinformatik allgemein bekannt. Zu den geeigneten Verfahren zählen beispielsweise die in der Bioinformatik allgemein bekannten Dynamic Programming - Algorithmen, z.B. der sogenannte Needleman-Wunsch-Algorithmus für globales Alignment und der Smith-Waterman-

Algorithmus für lokales Alignment. Insoweit wird beispielsweise auf Michael S. Waterman, Introduction to Computational Biology, London, New York 2000, insbesondere S. 207 bis 209 oder Dan Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge, 1999, insbesondere S. 215 bis 235, verwiesen.

Die Erfindung kann insbesondere vorsehen, daß jede Wiederholung eines partiellen Sequenzabschnittes in einem anderen Teil der Testsequenz oder eines vorgegebenen Bereichs der Testsequenz mit einem bestimmten Gewicht gewichtet wird, welches ein Maß für den Grad der Übereinstimmung und/oder die Größe der zueinander ähnlichen Abschnitte darstellt, und daß die Gewichte der einzelnen Wiederholungen zur Ermittlung des Kriteriumsgewichts betreffend die Wiederholungen bzw. invers komplementären Wiederholungen addiert werden. Es kann ebenfalls vorgesehen sein, daß die Gewichte der einzelnen Wiederholungen mit einem vorgegebenen Exponenten, dessen Wert vorzugsweise zwischen 1 und 2 liegt, potenziert werden und anschließend die Summation zur Ermittlung des Kriteriumsgewichts betreffend die Wiederholungen bzw. invers komplementäre Wiederholungen durchgeführt wird. Dabei kann vorgesehen sein, daß Wiederholungen unterhalb einer bestimmten Länge und/oder Wiederholungen, deren Gewichtsanteil unterhalb einer gewissen Schwelle liegt, nicht berücksichtigt werden. Die Erfindung kann vorsehen, daß zur Berechnung des entsprechenden Kriteriumsgewichts nur die Wiederholungen oder invers komplementären Wiederholungen eines partiellen Sequenzabschnitts berücksichtigt werden, der in einem vorgegebenen Teilbereich der Testsequenz (Testbereich), z.B. an dessen Ende und/oder in einem Variationsfenster liegt. Beispielsweise kann vorgesehen sein, daß nur die letzten 36 Basen der Testsequenz daraufhin überprüft werden, ob ein bestimmter Sequenzabschnitt innerhalb dieser 36 Basen mit einem anderen Sequenzabschnitt der gesamten Testsequenz oder der gesamten invers komplementären Testsequenz übereinstimmt.

Die Erfindung kann vorsehen, daß bei den Kriteriumsgewichten betreffend Wiederholungen, invers komplementäre Wiederholungen und/oder DNS-Motive nur der oder die M Abschnitte der Testsequenz berücksichtigt werden, welche den größten bzw. betragsmäßig größten Bei-

trag zu dem Kriteriumsgewicht liefern, wobei M eine natürliche Zahl, vorzugsweise zwischen 1 und 10, ist.

Gemäß einer Ausführungsform der Erfindung kann vorgesehen sein, daß eine Matrix generiert wird, deren Spaltenzahl der Anzahl der Positionen des Bereichs der Testsequenz (Testbereich) entspricht, der auf Wiederholungen in anderen Bereichen überprüft werden soll, und dessen Zeilenzahl der Anzahl der Positionen des Bereichs der Testsequenz entspricht, mit dem verglichen werden soll (Vergleichsbereich). Sowohl der Testbereich als auch der Vergleichsbereich können die gesamte Testsequenz umfassen.

Die Erfindung kann weiterhin vorsehen, daß die gesamte Gewichtsfunktion GesScore sich wie folgt bestimmt:

$$\text{GesScore} = \text{CUScore} - \text{GCScore} - \text{REPScore} - \text{SiteScore},$$

wobei CUScore das Kriteriumsgewicht für die Codon Usage ist, GCScore das Kriteriumsgewicht für den GC-Gehalt ist, REPScore das Kriteriumsgewicht für Wiederholungen und invers komplementäre Wiederholungen von gleichen oder ähnlichen Sequenzabschnitten ist und SiteScore das Kriteriumsgewicht für das Auftreten von unerwünschten bzw. geforderten Motiven ist.

Das Gewicht REPScore kann gemäß einer Ausführungsform der Erfindung aus einer Summe von zwei Bestandteilen bestehen, von denen der erste das Kriteriumsgewicht für die Wiederholung von gleichen oder ähnlichen Sequenzabschnitten in der Testsequenz selbst bzw. eines Teilbereichs davon angibt und der zweite Bestandteil das Kriteriumsgewicht für invers komplementäre Wiederholungen von gleichen oder ähnlichen Sequenzabschnitten in der Testsequenz oder eines Teilbereichs davon angibt.

Wenn die Gütefunktion sich aus Anteilen mehrerer Testkriterien zusammensetzt, insbesondere dann, wenn die Gütefunktion aus einer Linearkombination von Kriteriumsgewichten be-

steht, muß in einem Iterationsschritt eine Testsequenz nicht notwendigerweise nach allen Kriterien bewertet werden. Vielmehr kann die Bewertung bereits dann abgebrochen werden, wenn absehbar ist, daß der Wert der Gütefunktion geringer oder, allgemeiner gesprochen, weniger optimal, als der Wert der Gütefunktion einer bereits bewerteten Testsequenz ist. Bei den vorangehend beschriebenen Ausführungsformen gehen die meisten Kriterien, wie die Kriteriumsgewichte für repetitive Elemente, auszuschließende Motive usw., negativ in die Gütefunktion ein. Wenn nach Berechnung der Kriteriumsgewichte, welche positiv in die Gütefunktion eingehen und ggf. einem Teil der Kriteriumsgewichte, welche negativ in die Gütefunktion eingehen, sich bei der Aufsummation entsprechend der durch die Gütefunktion definierten Linearkombination der entsprechenden bereits berechneten Kriteriumsgewichte einen Wert ergibt, der kleiner ist als ein bereits berechneter Wert der vollständigen Gütefunktion für eine andere Testsequenz, kann die aktuell bewertete Testsequenz bereits ausgeschieden werden. Ebenso kann zum Beispiel dann, wenn ein Kriteriumsgewicht betragsmäßig wesentlich größer ist als alle anderen Gewichte, häufig die Bewertung bereits nach der Ermittlung des entsprechenden Kriteriumsgewichts abgebrochen werden. Wenn beispielsweise in einer ersten Testsequenz ein unerwünschtes Motiv nicht aufgetaucht ist und in einer zweiten Testsequenz das unerwünschte Motiv auftaucht, kann die zweite Testsequenz sofort ausgeschlossen werden, da das Kriteriumsgewicht für die Motivsuche so groß ist, daß es nicht durch andere Kriteriumsgewichte kompensiert werden kann.

Insbesondere kann die Erfindung vorsehen, daß bei Ausführungsformen, bei denen die Gütefunktion iterativ berechnet werden kann, zumindest bei einer Iteration eine obere (bzw. bei Optimierung auf das Minimum der Gütefunktion untere) Grenze bestimmt wird, unterhalb (bzw. oberhalb) derer der Wert der vollständigen Gütefunktion liegt, und die Iteration der Gütefunktion abgebrochen wird, wenn dieser Wert unter (bzw. über) dem Wert der vollständigen Gütefunktion für eine Testsequenz liegt, der vorangehend ermittelt wurde.

Die Erfindung kann in diesen Fällen vorsehen, daß im weiteren Verfahren für diese Testsequenz als Wert der Gütefunktion die besagte obere bzw. untere Grenze, falls erforderlich, verwendet wird und/oder daß die entsprechende Testsequenz in dem Algorithmus ausgeschie-

den wird, etwa dadurch, daß die Variable für die optimierte Testsequenz mit einer vorangehend aufgefundenen Testsequenz besetzt bleibt, bei der die Gütefunktion einen höheren Wert als die oben genannte Grenze, und der Algorithmus zu der Bewertung der nächsten Testsequenz übergeht. Die Erfindung kann dabei, insbesondere wenn die Gütefunktion eine Linearkombination von Kriteriumsgewichten ist, vorsehen, daß in den ersten Iterationen derjenige Beitrag oder diejenigen Beiträge berechnet werden, deren höchster Wert bzw. deren minimaler Wert den höchsten Absolutbetrag besitzt.

Die Erfindung kann vorsehen, daß bei einer Gütefunktion, die auf ihr Maximum optimiert wird und die durch eine Linearkombination von Kriteriumsgewichten gebildet wird, zunächst die positiven Anteile der Linearkombination berechnet werden und die Iteration abgebrochen wird, wenn in einer Iteration nach der Berechnung aller positiven Kriteriumsgewichte der Wert der Gütefunktion in dieser Iteration kleiner ist als der Wert der vollständigen Gütefunktion für eine andere Testsequenz.

Die Erfindung kann auch vorsehen, daß eine Iteration der Gütefunktion abgebrochen wird, wenn in einer Iteration festgestellt wird, daß die Summe aus dem in dieser Iteration berechneten Wert der Gütefunktion und dem Höchstwert des Beitrags der noch nicht berechneten Kriteriumsgewichte unterhalb des Werts der vollständigen Gütefunktion einer anderen Testsequenz liegt.

Das erfindungsgemäße Verfahren kann den Schritt des Synthetisierens der optimierten Nucleotidsequenz umfassen.

Dabei kann vorgesehen sein, daß der Schritt des Synthetisierens der optimierten Nucleotidsequenz in einer Vorrichtung zum automatischen Synthetisieren von Nucleotidsequenzen, zum Beispiel in einem Oligonucleotidsynthesizer, stattfindet, welcher von dem Rechner angesteuert wird, der die Nucleotidsequenz optimiert.

Die Erfindung kann insbesondere vorsehen, daß der Rechner, sobald der Optimierungsprozeß abgeschlossen ist, die ermittelten Daten über die optimale Nucleotidsequenz an einen Oligonucleotidsynthesizer weitergibt und diesen veranlaßt, die Synthese der optimierten Nucleotidsequenz durchzuführen.

Diese Nucleotidsequenz kann dann, wie gewünscht, hergestellt werden. Zur Expression des Proteins wird die entsprechende Nucleotidsequenz in Wirtszellen eines Wirtsorganismus eingebracht, auf welchen sie optimiert ist und welcher dann letztendlich das Protein erzeugt.

Die Erfindung stellt auch eine Vorrichtung zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins zur Verfügung, welche eine Recheneinrichtung aufweist, welche umfaßt:

- eine Einrichtung zum Generieren einer ersten Testsequenz von n Codons, welche n aufeinanderfolgenden Aminosäuren in der Proteinsequenz entsprechen, wobei n eine natürliche Zahl kleiner oder gleich N , der Zahl der Aminosäuren der Proteinsequenz ist,
- eine Einrichtung zum Festlegen von n Optimierungspositionen in der Testsequenz, welche der Position von m Codons entsprechen, an denen die Besetzung mit einem Codon, bezogen auf die Testsequenz, optimiert werden soll, wobei $m \leq n$ und $m < M$ ist,
- eine Einrichtung zum Generieren einer oder mehrerer weiterer Testsequenzen aus der ersten Testsequenz, indem an einer oder mehreren der m Optimierungspositionen ein Codon der ersten Testsequenz durch ein anderes Codon ersetzt wird, welches dieselbe Aminosäure exprimiert,
- eine Einrichtung zum Bewerten jeder der Testsequenzen mit einer Gütefunktion und zum Ermitteln der hinsichtlich der Gütefunktion optimalen Testsequenz,
- eine Einrichtung zum Festlegen von p Codons der optimalen Testsequenz, welche sich an einem der m Optimierungspositionen befinden, als Ergebniscodons, welche die Codons der optimierten Nucleotidsequenz an den Positionen bilden, die den Positionen

der besagten p Codons in der Testsequenz entsprechen, wobei p eine natürliche Zahl und $p \leq m$ ist,

eine Einrichtung zum Iterieren der Schritte des Generierens mehrerer Testfunktionen, der Bewertung der Testsequenzen und des Festlegens von Ergebniscodons, vorzugsweise bis alle Codons der optimierten Nucleotidsequenz festgelegt worden sind, wobei in jedem Iterationsschritt die Testsequenz an den Positionen, welche Positionen von festgelegten Ergebniscodons in der optimierten Nucleotidsequenz entsprechen, das entsprechende Ergebniscodon enthält und die Optimierungspositionen von Positionen von Ergebniscodons verschieden sind.

Die vorangehend genannten Einrichtungen müssen nicht verschieden sein, sondern können insbesondere durch eine einzige Vorrichtung realisiert werden, welche die Funktionen der vorangehend genannten Einrichtungen realisiert.

Die erfindungsgemäße Vorrichtung kann allgemein eine Einrichtung zum Durchführen der Schritte der vorangehend beschriebenen Verfahren aufweisen.

Die erfindungsgemäße Vorrichtung kann einen Oligonucleotidsynthesizer aufweisen, welcher von dem Rechner so angesteuert wird, daß er die optimierte Nucleotidsequenz synthetisiert.

Bei dieser Ausführungsform der Erfindung kann entweder automatisch oder durch einen entsprechenden Befehl des Benutzers die optimierte Nucleotidsequenz synthetisiert werden, ohne daß Datentransfers, Einstellung von Parametern und dergleichen nötig sind.

Die Erfindung stellt auch ein Computerprogramm zur Verfügung, welches von einem Computer ausführbaren Programmcode enthält, der, wenn er auf einem Computer ausgeführt wird, den Computer veranlaßt, ein erfindungsgemäßes Verfahren durchzuführen.

Dabei kann der Programmcode, wenn er auf einem Computer ausgeführt wird, eine Vorrichtung zum automatischen Synthetisieren von Nucleotidsequenzen veranlassen, die optimierte Nucleotidsequenz herzustellen.

Die Erfindung stellt auch einen computerlesbaren Datenträger zur Verfügung, auf welchem in computerlesbarer Form ein erfindungsgemäßes Programm gespeichert ist.

Die Erfindung stellt weiterhin eine nach einem erfindungsgemäßen Verfahren hergestellte oder herstellbare Nukleinsäure und einen Vektor, der eine solche Nukleinsäure enthält, zur Verfügung. Die Erfindung stellt weiterhin eine Zelle, die einen solchen Vektor oder eine solche Nukleinsäure enthält, zur Verfügung sowie einen nicht-menschlichen Organismus bzw. ein nicht-menschliches Lebewesen, das eine solche Zelle enthält, wobei ein solches nicht-menschliches Lebewesen auch ein Säugetier sein könnte.

Während bei statistischen Verfahren keinerlei Korrelation zwischen einer Sequenz in einem vorangehenden Iterationsschritt und der Sequenz in einem nachfolgenden Iterationsschritt besteht, wird erfindungsgemäß in jedem Iterationsschritt zumindest ein Codon neu festgelegt. Da die Testsequenz nur auf einem Teil der Gesamtsequenz variiert wird, ist das Verfahren mit einem geringeren Aufwand durchführbar. Insbesondere ist es möglich, in dem Variationsbereich sämtliche möglichen Kombinationen von Codons zu evaluieren. Die Erfindung macht sich in vorteilhafter Weise den Umstand zunutze, daß langreichweitige Korrelationen innerhalb einer Nucleotidsequenz von untergeordneter Bedeutung sind, d.h. daß zur Erzielung eines akzeptablen Optimierungsergebnisses die Codons an einer Position weitgehend unabhängig von den Codons an einer weiter entfernten Position variiert werden können.

Das erfindungsgemäße Verfahren eröffnet in größerem Umfang als die bisherigen Verfahren die Möglichkeit, relevante biologische Kriterien in die Bewertung einer Testsequenz einzubeziehen. Beispielsweise können mit dem erfindungsgemäßen Verfahren erwünschte oder unerwünschte Motive in der synthetischen Nukleotidsequenz berücksichtigt werden. Da bei einer Motivsuche bereits ein individuelles Codon dafür ausschlaggebend sein kann, ob ein be-

stimmtes Motiv vorhanden ist oder nicht, werden rein stochastische Verfahren nicht oder nur mit einer sehr geringen Wahrscheinlichkeit optimierte Sequenzen liefern, welche ein gefordertes Motiv enthalten. Bei dem erfindungsgemäßen Verfahren ist dies jedoch deswegen möglich, da über einem Teilbereich der Sequenz sämtliche Codonkombinationen durchgetestet werden. Gegebenenfalls kann man, um das Vorhandensein bzw. Nichtvorhandensein eines bestimmten Sequenzmotivs zu gewährleisten, die Anzahl m der Optimierungspositionen so groß machen, daß diese größer ist als die Zahl der Codonpositionen (oder die Anzahl der Basenpositionen, geteilt durch 3) des entsprechenden Motivs. Wenn die m Optimierungspositionen zusammenhängend sind, ist damit gewährleistet, daß das Auftauchen eines bestimmten Sequenzmotivs zuverlässig erfaßt und das entsprechende Motiv in der Sequenz gewährleistet bzw. aus dieser ausgeschlossen werden kann. Die numerische Berechnung der Gütefunktion hat besondere Vorteile bei der Verwendung von Gewichtsmatrix-Scans. Da hierbei den verschiedenen Basen einer Erkennungssequenz eine unterschiedlich starke Bedeutung für die Erkennung bzw. die biologische Aktivität zugeordnet werden kann, kann bei dem erfindungsgemäßen Verfahren, bei dem über einen Teilbereich der Sequenz alle möglichen Codonkombinationen durchgetestet werden, die Sequenz gefunden werden, die zum Beispiel ein DNA-Motiv durch Eliminierung der für die Aktivität wichtigsten Basen am effektivsten ausschaltet bzw. es kann eine optimierte Kompromißlösung unter Einbeziehung anderer Kriterien gefunden werden.

Die Erfindung ist grundsätzlich nicht auf einen bestimmten Organismus beschränkt. Organismen, für welche eine Optimierung einer Nukleotidsequenz zur Expression eines Proteins mit dem erfindungsgemäßen Verfahren von besonderem Interesse ist, sind z.B. Organismen aus der folgenden Gruppe:

- Viren, insbesondere Vaccinia-Viren,
- Prokaryonten, insbesondere Escherichia coli, Caulobacter crescentus, Bacillus subtilis, Mycobacterium spec.,
- Hefen, insbesondere Saccharomyces cerevisiae, Schizosaccharomyces pombe, Pichia pastoris, Pichia angusta,
- Insekten, insbesondere Sprodoptera frugiperda, Drosophila spec,

- Säuger, insbesondere *Homo sapiens*, *Macaca mulata*, *Mus musculus*, *Bos taurus*, *Capra hircus*, *Ovis aries*, *Oryctolagus cuniculus*, *Rattus norvegicus*, chinese hamster ovary,
- monokotyle Pflanzen, insbesondere *Oryza sativa*, *Zea mays*, *Triticum aestivum*.
- dikotyle Pflanzen, insbesondere *Glycin max*, *Gossypium hirsutum*, *Nicotiana tabacum*, *Arabidopsis thaliana*, *Solanum tuberosum*.

Proteine, für die eine optimierte Nucleotidsequenz mit dem erfindungsgemäßen Verfahren generiert werden kann, sind zum Beispiel:

- Enzyme, insbesondere Polymerasen, Endonukleasen, Ligasen, Lipasen, Proteasen, Kinasen, Phosphatasen, Topoisomerasen,
- Cytokine, Chemokine, Transkriptionsfaktoren, Oncogene,
- Proteine aus thermophilen Organismen, aus cryophilen Organismen, aus halophilen Organismen, aus acidophilen Organismen, aus basophilen Organismen,
- Proteine mit repetitiven Sequenzelementen, insbesondere strukturgebende Proteine,
- Humane Antigene, insbesondere Tumorantigene, Tumormarker, Autoimmunantigene, diagnostische Marker,
- Virale Antigene, insbesondere von HAV, HBV, HCV, HIV, SIV, FIV, HPV, Rinöviren, Influenzaviren, Herpesviren, Poliomaviren, Hendra Virus, Dengue Virus, AAV, Adenoviren, HTLV, RSV,
- Antigene von Protozoen und/oder parasitären Erregern, insbesondere Erreger von Malaria, Leishmania, Trypanosoma, Toxoplasmen, Amöba,
- Antigene von bakteriellen Erregern oder Pathogene, insbesondere von den Genera Chlamydia, Staphylococci, Klebsiella, Streptococcus, Salmonella, Listeria, Borrelia, Escherichia coli,
- Antigene von Organismen der Sicherheitstufe L4, insbesondere Bacillus anthracis, Ebola-Virus, Marburg-Virus, Pockenviren.

Die vorangehende Aufzählung von Organismen bzw. Proteinen, für welche die Erfindung Anwendung findet, ist in keiner Weise einschränkend und lediglich als Beispiel zur besseren Veranschaulichung gedacht.

Weitere Merkmale und Vorteile der Erfindung ergeben sich aus der nachfolgenden Beschreibung von Ausführungsbeispielen der Erfindung anhand der beigefügten Zeichnungen.

Figur 1a, 1b zeigen ein Flußdiagramm eines Ausführungsbeispiels des Verfahrens der Erfindung,

Figur 2 illustriert das Verhältnis von Testsequenz, optimierter DNS-Sequenz, Kombinations-DNS-Sequenz und Aminosäuresequenz für ein Ausführungsbeispiel der Erfindung,

Figur 3 zeigt die Bereiche für die Bestimmung der Sequenzwiederholung,

Figur 4a und 4b zeigen schematisch ein Schema für die Bestimmung von Sequenzwiederholungen,

Figur 5a zeigt die Codon usage bei einer ausschließlichen Optimierung auf die Codon usage,

Figur 5b zeigt den GC-Gehalt bei einer ausschließlichen Optimierung auf die Codon usage,

Figur 6a zeigt die Codon usage bei Verwendung einer ersten Gütefunktion,

Figur 6b zeigt den GC-Gehalt bei Verwendung einer ersten Gütefunktion,

Figur 7a zeigt die Codon usage bei Verwendung einer zweiten Gütefunktion,

Figur 7b zeigt den GC-Gehalt bei Verwendung einer zweiten Gütefunktion,

Figur 8a zeigt die Codon usage bei Verwendung einer dritten Gütefunktion,

Figur 8b zeigt den GC-Gehalt bei Verwendung einer dritten Gütefunktion.

Gemäß einer bevorzugten Ausführungsform der Erfindung wird in einer Iteration die Wahl des Codons für die i -te Aminosäure einer Aminosäuresequenz der Länge N betrachtet. Dazu werden sämtliche möglichen Codonkombinationen der verfügbaren Codons für die Aminosäuren an den Positionen i bis $i + m - 1$ gebildet. Diese Positionen bilden ein Variationsfenster und legen die Optimierungspositionen fest, auf denen die Sequenz variiert werden soll. Jede Kombination von Codons auf diesem Variationsfenster resultiert in einer DNS-Sequenz mit $3m$ Basen, die im folgenden Kombinations-DNS-Sequenz (KDS) genannt wird. In jedem Iterationsschritt wird zu jeder KDS eine Testsequenz gebildet, welche die KDS an ihrem Ende enthält. Im ersten Iterationsschritt bestehen die Testsequenzen nur aus den Kombinations-DNS-Sequenzen. Die Testsequenzen werden mit einer nachfolgend näher beschriebenen Gütefunktion gewichtet und das erste Codon derjenigen KDS, welche den maximalen Wert der Gütefunktion aufweist, wird für alle weiteren Iterationen als Codon der optimierten Nucleotidsequenz (Ergebniscodon) beibehalten. Dies bedeutet, daß dann, wenn in einer Iteration das i -te Codon festgelegt wurde, jede der Testsequenzen in der nächsten Iteration dieses Codon an der Position i enthält und an den Positionen $i + 1$ bis $i + m$ die Codons der verschiedenen Kombinations-DNS-Sequenzen. Bei der j -ten Iteration bestehen also alle Testsequenzen an den Positionen 1 bis $j - 1$ aus den in den vorangehenden Iterationen als optimal aufgefundenen Codons, während die Codons an den Positionen j bis $j + m - 1$ variiert werden. Die Güte der DNS-Sequenz läßt sich für jedes individuelle Testkriterium als Kriteriumsgewicht (Einzelscore) ausdrücken. Durch Addition der nach benutzerdefinierten Vorgaben gewichteten Kriteriumsgewichte wird ein Gesamtgewicht (Gesamtscore) gebildet, welches den Wert der

Gütefunktion für die gesamte Testsequenz angibt. Wenn $j = N - m + 1$ ist, ist die optimale Testsequenz gleichzeitig die optimierte Nucleotidsequenz nach dem Verfahren der Erfindung. Daher werden in diesem (letzten) Schritt sämtliche Codons der optimal KDS als Codons der optimierten Nucleotidsequenz festgelegt.

Der vorangehend beschriebene Ablauf ist schematisch in Figur 1 illustriert. Der Algorithmus beginnt bei der ersten Aminosäure ($i=1$). Es wird nun eine erste KDS der Codons für die Aminosäuren i bis $i + m - 1$ gebildet (bei der ersten Iteration sind dies die Aminosäuren 1 bis m). Diese KDS wird mit der bereits optimierten DNS-Sequenz zu einer Testsequenz zusammengefügt. Im ersten Schritt besteht die optimierte DNS-Sequenz aus 0 Elementen. Daher besteht die Testsequenz bei der ersten Iteration nur aus der zuvor gebildeten (ersten) KDS.

Die Testsequenz wird nun nach benutzerdefinierten Kriterien evaluiert. Der Wert einer Gütefunktion wird berechnet, indem Kriteriumsgewichte für verschiedene Bewertungskriterien berechnet und in einer Bewertungsfunktion verrechnet werden. Wenn der Wert der Gütefunktion besser als ein gespeicherter Wert der Gütefunktion ist, wird der neue Wert der Gütefunktion gespeichert. Gleichzeitig wird auch das erste Codon der zugehörigen KDS, welches die Aminosäure i repräsentiert, gespeichert. Wenn der Wert der Gütefunktion schlechter als der gespeicherte Wert ist, erfolgt keine Maßnahme. Im nächsten Schritt wird überprüft, ob alle möglichen KDS gebildet worden sind. Ist dies nicht der Fall, wird die nächstmögliche KDS gebildet und mit der bereits optimierten DNS-Sequenz zu einer neuen Testsequenz zusammengefügt. Die Schritte des Evaluierens, des Bestimmens einer Gütefunktion und des Vergleichs des Wertes der Gütefunktion mit einem gespeicherten Wert wiederholen sich dann. Sind dagegen alle möglichen KDS gebildet worden, wird, sofern $i \neq N - m + 1$ ist, das gespeicherte Codon an die bereits gebildete optimierte DNS-Sequenz an der Position i angefügt. Bei der ersten Iteration wird die optimierte DNS-Sequenz dadurch gebildet, daß das gespeicherte Codon auf die Position 1 der optimierten DNS-Sequenz gesetzt wird. Der Prozeß wiederholt sich dann für die nächste Aminosäure ($i + 1$). Ist dagegen $i = N - m + 1$, wird die gesamte KDS der optimalen Testsequenz an die bereits gebildete optimierte DNS-Sequenz an-

gehängt, da sie bereits hinsichtlich der Bewertungskriterien optimiert ist. Es folgt dann die Ausgabe der optimierten Sequenz.

Das Verhältnis der verschiedenen Bereiche ist diagrammatisch in Figur 2 dargestellt. Man erkennt die Kombinations-DNS-Sequenz und den Bereich der bereits festgelegten optimierten DNS-Sequenz.

Der Parameter m kann in weiten Bereichen variiert werden, wobei im Sinne einer bestmöglichen Optimierung eine möglichst hohe Zahl von variierten Codons angestrebt wird. Mit den derzeit verfügbaren Rechnern läßt sich mit einer Größe des Variationsfensters von $m = 5$ bis $m = 10$ in einer akzeptablen Zeit ein sinnvolles Optimierungsergebnis erreichen.

Neben der individuellen Gewichtung der Kriteriumsgewichte können sowohl das Gesamtgewicht als auch die Kriteriumsgewichte durch geeignete mathematische Funktionen definiert sein, die gegenüber den einfachen Relationen, wie Differenz oder Proportion, modifiziert sind, z.B. durch abschnittsweise definierte Funktionen, welche einen Schwellenwert definieren, oder nichtlineare Funktionen. Ersteres ist beispielsweise bei der Bewertung von Wiederholungen oder invers komplementären Wiederholungen sinnvoll, die erst ab einer bestimmten Größe berücksichtigt werden sollen. Letzteres ist z.B. bei der Bewertung der Codon usage oder des CG-Gehalts sinnvoll.

Nachfolgend werden verschiedene beispielhafte Gewichtungskriterien erläutert, die erfindungsgemäß verwendet werden können, ohne daß die Erfindung auf diese Kriterien bzw. die nachfolgend beschriebenen Gewichtungsfunktionen beschränkt wäre.

Die Anpassung der Codon usage des synthetischen Gens an die Codonusage des Wirtsorganismus ist eines der wichtigsten Kriterien bei der Optimierung. Hierbei muß die unterschiedliche Degeneriertheit der verschiedenen Codons (einfach bis sechsfach) berücksichtigt werden. Hierfür geeignete Größen sind z.B. die RSCU (relative synonymous codon usage) oder

relative Häufigkeiten (Relative Adaptiveness), die auf die Häufigkeit des am meisten von dem Organismus genutzten Codons normiert sind (das am meisten genutzte Codon hat also die „Codon usage“ 1), vgl. P. M. Sharp, W. H. Li, Nucleic Acid Research 15 (1987), 1281 bis 1295.

Zur Bewertung einer Testsequenz wird bei einer Ausführungsform der Erfindung die durchschnittliche Codon usage auf dem Variationsfenster verwendet.

Bei der Bewertung des GC-Gehalts ist eine möglichst geringe Abweichung des durchschnittlichen GC-Gehaltes von dem vorgegebenen gewünschten GC-Gehalt erforderlich. Weiterhin ist es anzustreben, Schwankungen des GC-Gehaltes über dem Verlauf der Sequenz gering zu halten.

Zur Evaluierung einer Testsequenz wird der durchschnittliche prozentuale GC-Gehalt desjenigen Bereichs der Testsequenz ermittelt, der die KDS und vor dem Beginn der KDS liegende Basen umfaßt, deren Anzahl b vorzugsweise zwischen 20 und 30 Basen liegt. Das Kriteriumsgewicht wird aus dem Absolutwert der Differenz zwischen dem gewünschten GC-Gehalt und dem ermittelten GC-Gehalt für die Testsequenz ermittelt, wobei dieser Absolutwert als Argument in eine nichtlineare Funktion, z.B. in eine Exponentialfunktion eingehen kann.

Wenn das Variationsfenster eine Breite von mehr als 10 Codonpositionen hat, können Schwankungen des GC-Gehalts innerhalb der KDS von Bedeutung sein. In diesen Fällen wird, wie vorangehend erläutert, der GC-Gehalt für jede Basenposition auf einem Fenster ermittelt, das bezüglich der Basenposition in einer bestimmten Weise ausgerichtet ist und eine bestimmte Anzahl, zum Beispiel 40 Basen, umfassen kann, und die Absolutwerte der Differenz zwischen dem gewünschten GC-Gehalt und dem für jede Basenposition ermittelten „lokalen“ GC-Gehalt werden aufsummiert. Teilt man die Summe durch die Anzahl der ermittelten Einzelwerte, so erhält man als Kriteriumsgewicht die durchschnittliche Abweichung von dem gewünschten GC-Gehalt. Bei dem vorangehend beschriebenen Vorgehen kann die Lage

des Fensters so definiert sein, daß die besagte Basenposition zum Beispiel am Rand oder im Zentrum des Fensters liegt. Alternativ kann auch als Kriterium der Absolutbetrag der Differenz zwischen dem tatsächlichen GC-Gehalt in der Testsequenz oder auf einem Teilbereich davon zu dem gewünschten GC-Gehalt oder der Absolutbetrag der Differenz zwischen dem Mittelwert des vorangehend erwähnten „lokalen“ GC-Gehalts über die Testsquenz oder einem Teil davon und dem gewünschten GC-Gehalt als Kriterium verwendet werden. In einer weiteren Abwandlung kann auch vorgesehen sein, daß das entsprechende Kriteriumsgewicht proportional zu dem Quadrat der Differenz zwischen dem tatsächlichen GC-Gehalt und dem gewünschten GC-Gehalt, dem Quadrat der Differenz zwischen dem über die Basenpositionen gemittelten GC-Gehalt und dem gewünschten GC-Gehalt bzw. der Mittelwert des Quadrats der Differenzen zwischen dem lokalen GC-Gehalt und dem gewünschten GC-Gehalt als Kriterium verwendet werden. Das Kriteriumsgewicht für den GC-Gehalt hat das entgegengesetzte Vorzeichen wie das Kriteriumsgewicht für die Codon usage.

Lokale Erkennungssequenzen bzw. biophysikalische Charakteristika spielen in der Zell- und Molekularbiologie eine entscheidende Rolle. Eine unbeabsichtigte Generierung entsprechender Motive innerhalb der Sequenz des synthetisierten Gens kann unerwünschte Wirkungen haben. Zum Beispiel kann die Expression stark reduziert oder ganz unterdrückt werden; es kann auch eine für den Wirtsorganismus toxische Wirkung entstehen. Bei der Optimierung der Nucleotidsequenz ist es daher wünschenswert, die unbeabsichtigte Generierung solcher Motive auszuschließen. Im einfachsten Fall läßt sich die Erkennungssequenz durch eine gut charakterisierte Consensussequenz (z.B. Restriktionsenzym-Erkennungssequenz) unter Verwendung entsprechender IUPAC-Basensymbole darstellen. Führt man eine einfache Regular-Expressionssuche innerhalb der Testsequenz durch, so erhält man für die Berechnung des entsprechenden Gewichts die Anzahl der aufgefundenen Positionen. Läßt man eine bestimmte Anzahl von Fehlstellen (mismatches) zu, muß die Anzahl der Fehlstellen bei einer erkannten Übereinstimmung bei der Ermittlung der Gewichtsfunktion berücksichtigt werden, zum Beispiel derart, daß das lokale Gewicht für eine Basenposition umgekehrt proportional zu der Anzahl der Basen ist, die einem IUPAC-Consensusymbol zugeordnet sind. In vielen Fällen ist die Consensussequenz jedoch nicht ausreichend eindeutig (vgl. zum Beispiel K. Quandt

u.a., Nucleic Acid Research 23 (1995), 4878). In solchen Fällen kann man auf eine Matrizendarstellung der Motive zurückgreifen oder andere Erkennungsmethoden, z.B. mittels neuraler Netze, verwenden.

Bei der bevorzugten Ausführungsform der Erfindung wird für jedes aufgefundene Motiv ein Wert zwischen 0 und 1 bestimmt, der im Idealfall die Bindungsaffinität der gefundenen (potentiellen) Stelle bzw. deren biologische Aktivität oder auch deren Erkennungssicherheit widerspiegelt. Für die Berechnung des Kriteriumsgewichts für DNS-Motive wird dieser Wert mit einem geeigneten Gewichtungsfaktor multipliziert und die Einzelwerte für jede aufgefundene Übereinstimmung werden addiert.

Das Gewicht für unerwünschte Motive geht mit dem umgekehrten Vorzeichen wie dasjenige für die Codon usage in die Gesamtgütefunktion ein.

In der gleichen Weise kann in die Gewichtung das Vorhandensein bestimmter erwünschter DNS-Motive, z.B. RE-Schnittsequenzen, bestimmte Enhancersequenzen oder immunstimulatorische bzw. immunsupprimierende CpG-Motive einbezogen werden. Das Gewicht für erwünschte DNS-Motive geht mit dem gleichen Vorzeichen wie das Gewicht für die Codon usage in die Gesamtbewertung ein.

Stark repetitive Sequenzabschnitte können zum Beispiel zu einer geringen genetischen Stabilität führen. Die Synthese repetitiver Abschnitte ist auch wegen der Gefahr von Fehlhybridisierung deutlich erschwert. Gemäß der bevorzugten Ausführungsform der Erfindung geht daher in die Bewertung einer Testsequenz ein, ob diese an unterschiedlichen Stellen identische oder einander ähnliche Sequenzabschnitte enthält. Das Vorhandensein entsprechender Abschnitte kann beispielsweise mit Hilfe einer Variante eines Dynamic Programming - Algorithmus zur Generierung eines lokalen Alignments der einander ähnlichen Sequenzabschnitte festgestellt werden. Wichtig bei dieser Ausführungsform der Erfindung ist, daß der verwendete Algorithmus einen Wert generiert, welcher geeignet ist, den Grad der Übereinstimmung

und/oder die Länge der einander ähnlichen Sequenzabschnitte quantitativ zu beschreiben (Alignmentgewicht). Hinsichtlich weiterer Einzelheiten betreffend einen möglichen Algorithmus wird auf die oben genannten Lehrbücher von Gusfield oder Waterman bzw. M. S. Waterman, M. Eggert, J. Mol. Biology, (1987) 197, 723 bis 728, verwiesen.

Zur Berechnung des Kriteriumsgewichts hinsichtlich der repetitiven Elemente summiert man die Einzelgewichte aller lokalen Alignments, bei denen das Alignmentgewicht einen bestimmten Schwellenwert übersteigt. Die Addition dieser Einzelgewichte ergibt das Kriteriumsgewicht, welches die Repetitivität der Testsequenz charakterisiert.

Gemäß einer Abwandlung der vorangehend beschriebenen Ausführungsform wird nur der eine Bereich am Ende der Testsequenz, welcher das Variationsfenster sowie eine gewisse Anzahl weiterer Basen, z.B. 20 bis 30, umfaßt, daraufhin überprüft, ob ein Teilabschnitt der Testsequenz in diesem Bereich einer anderen Stelle der Testsequenz in gleicher oder ähnlicher Weise vorkommt. Dies ist schematisch in Figur 3 dargestellt. Die durchgezogene Linie in der Mitte stellt die gesamte Testsequenz dar. Die obere Linie stellt die KDS dar, während der untere Bereich den Vergleichsbereich der Testsequenz darstellt, welcher mit der restlichen Testsequenz auf übereinstimmende Sequenzabschnitte überprüft wird. Die Überprüfung der Testsequenzen auf übereinstimmende oder ähnliche Abschnitte des Vergleichsbereichs (vgl. Figur 3) mit der Dynamic Programming-Matrixtechnik ist in Figur 4a und 4b illustriert. Figur 4a zeigt den Fall, daß ähnliche oder übereinstimmende Sequenzabschnitte A und B in dem Vergleichsbereich selbst vorhanden sind. Figur 4b zeigt den Fall, daß ein Sequenzabschnitt B in dem Vergleichsbereich mit einem Sequenzabschnitt A außerhalb des Vergleichsbereichs übereinstimmt oder diesem ähnlich ist.

Als Alternative zu der Summation von Einzelgewichten kann auch vorgesehen sein, daß nur dasjenige Alignment, das zu dem höchsten Einzelgewicht führt oder, allgemeiner, nur die Alignments mit den m größten Einzelgewichten, berücksichtigt werden.

Mit der vorangehend beschriebenen Gewichtung können sowohl ähnliche Sequenzen, die z.B. am Anfang und am Ende der Testsequenz vorhanden sind, als auch sogenannte Tandem-Repeats, bei denen sich die ähnlichen Bereiche beide am Ende der Sequenz befinden, erfaßt werden.

Invers komplementäre Wiederholungen können in der gleichen Weise wie einfache Wiederholungen behandelt werden. Die potentielle Bildung von Sekundärstrukturen auf RNA-Ebene oder cruciformer Strukturen auf DNS-Ebene läßt sich an der Testsequenz durch das Vorhandensein solcher invers komplementärer Wiederholungen (inverse Repeats) erkennen. Cruciforme Strukturen auf DNS-Ebene können die Translation behindern und zu genetischer Instabilität führen. Man vermutet, daß die Bildung von Sekundärstrukturen auf RNA-Ebene sich negativ auf die Translationseffizienz auswirkt. Dabei sind insbesondere solche inverse Repeats von Bedeutung, die Haarnadelschleifen bzw. cruciforme Strukturen ausbilden. Fehlhybridisierungen oder Haarnadelschleifen können sich auch bei der Synthetisierung jener aus Oligonucleotiden negativ auswirken.

Die Überprüfung auf invers komplementäre Wiederholungen erfolgt vom Grundsatz her analog zur Überprüfung auf einfache Wiederholungen. Die Testsequenz bzw. der Vergleichsbereich der Testsequenz wird jedoch mit der invers komplementären Sequenz verglichen. In einer Fortbildung kann die thermodynamische Stabilität bei dem Vergleich („alignment“) berücksichtigt werden, im einfachsten Fall durch die Verwendung einer Scoring Matrix. Dabei wird z.B. ein Match CC bzw. GG aufgrund der stabileren Basenpaarung stärker gewichtet als eine Überweinstimmung TT oder AA. Entsprechend können auch Fehlstellen (mismatches) variabel gewichtet werden. Eine spezifischere Gewichtung kann dadurch erfolgen, daß Nearest-Neighbour-Parameter zur Berechnung der thermodynamischen Stabilität verwendet werden, was allerdings den Algorithmus komplexer macht. Hinsichtlich eines möglichen Algorithmus wird beispielsweise auf L. Kaderali, A. Schliep, Bioinformatics 18 (10) 2002, 1340 bis 1349 verwiesen.

Bei allen Bewertungskriterien kann die Erfindung vorsehen, daß die entsprechende Gewichtungsfunktion positionsabhängig ist. Beispielsweise kann die Generierung einer RE-Schnittsequenz an einer bestimmten Stelle stärker gewichtet werden oder Sekundärstrukturen können am 5'-Ende stärker gewichtet werden, da sie dort stärker inhibierend sind. Ebenso kann der Codonkontext, d.h. das oder die Vorgänger- bzw. Nachfolgerkodons, berücksichtigt werden. Weiterhin kann für bestimmte Codons, deren Verwendung an den Domänengrenzen eine Rolle bei der cotranslatorischen Proteinfaltung spielt, ein Beitrag zur Gütefunktion vorgesehen sein, der davon abhängt, ob dieses Codon näher an der Domänengrenze liegt oder nicht. Weitere Kriterien, die in die Gütefunktion eingehen können, sind z.B. biophysikalische Eigenschaften, wie die Steifigkeit oder die Krümmung der DNS-Sequenz. Je nach Anwendungsgebiet können auch Kriterien einfließen, die mit weiteren DNS-Sequenzen assoziiert sind. Beispielsweise ist im Bereich der DNS-Vakzinierung entscheidend, daß die für die Vakzinierung verwendeten Sequenzen keine signifikante Ähnlichkeit mit den pathogenen Elementen des natürlichen Virusgenoms aufweisen, um unerwünschte Rekombinationsereignisse sicher auszuschließen. In gleicher Weise sollten die für gentherapeutische Zwecke verwendeten Vektoren eine möglichst geringe Ähnlichkeit zu Sequenzen des menschlichen Genoms aufweisen, um einerseits homologe Rekombination in das menschliche Genom auszuschließen und andererseits ein selektives Abschalten von vitalen Genen in Transkriptom durch RNA-Interferenz-Phänomene (RNAI - Phänomene) zu vermeiden. Letzteres ist auch von allgemeiner Bedeutung bei der Herstellung von rekombinanten Zellfabriken und insbesondere bei transgenen Organismen.

Erfindungsgemäß können die verschiedenen Kriteriumsgewichte für verschiedene Kriterien unterschiedlich in die Gesamtgewichtsfunktion eingehen. Dabei ist der durch das entsprechende Kriterium maximal erreichbare Unterschied in dem Wert der Gütefunktion für die gebildete Testsequenz wichtig. Einen hohen Anteil an bestimmten Kriteriumsgewichten haben jedoch DNA-Basen, welche durch unterschiedliche KDS nicht geändert werden können, wie z.B. die in die Berechnung des durchschnittlichen GC-Gehalts miteinbezogenen Nucleotide vor der KDS und die innerhalb synonymen Codons unveränderlichen Nucleotide. Die individuelle Gewichtung eines Kriteriums gegenüber anderen Kriterien kann daher davon ab-

hängig gemacht werden, wie stark die Güte der Testsequenz von der Zielvorgabe abweicht. Es kann sinnvoll sein, die Kriteriumsgewichte zur weiteren Verarbeitung in mathematischen Funktionen zu Berechnung der Gütefunktion aufzuspalten in einen Teil, der den bei Verwendung unterschiedlicher KDS variablen Anteil eines Kriteriums bemißt und einen Teil, der die unveränderlichen Anteile bemißt.

Die vorangehend beschriebenen Ausführungsformen der Erfindung werden nachfolgend anhand zweier konkreter Beispiele weiter erläutert.

Beispiel 1

Zu der nachfolgend gezeigten (fiktiven) Aminosäuresequenz ASSeq1 soll die zugehörige optimale DNS-Sequenz ermittelt werden. Als Referenz dient eine konventionelle Rückübersetzung mit Optimierung auf optimale Codon-Usage.

ASSeq1:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
E	Q	F	I	I	K	N	M	F	I	I	K	N	A
GAA	CAG	TTT	ATT	ATT	AAA	AAC	ATG	TTT	ATT	ATT	AAA	AAC	GCG
GAG	CAA	TTC	ATC	ATC	AAG	AAT		TTC	ATC	ATC	AAG	AAT	GCC
			ATA	ATA					ATA	ATA			GCA
													GCT

Folgende Kriterien werden der Optimierung zugrunde gelegt:

- Die Codon usage soll auf die Codon usage von E. Coli K12 optimiert werden.
- Der GC – Gehalt soll möglichst nahe bei 50 % liegen.
- Repetitionen sollen möglichst ausgeschlossen werden
- Die Nla III Erkennungssequenz CATG soll ausgeschlossen werden

Als Bewertungsfunktion für die Codon usage wird folgende Funktion verwendet:

$$CUScore = \langle CU \rangle$$

wobei $\langle CU \rangle$ bei diesem Beispiel das arithmetische Mittel der Relative adaptiveness über den Codonpositionen der Testsequenz ist.

Zur Darstellung der Codon usage eines Codons wird zur besseren Vergleichbarkeit der Codongüte verschiedener Aminosäuren das jeweils beste Codon für eine bestimmte Aminosäure gleich 100 gesetzt und die schlechteren Codons entsprechend ihrem tabellierten prozentualen Anteil reskaliert. Ein CUScore von 100 bedeutet also, daß ausschließlich die für E. Coli K12 optimalen Codons verwendet werden.

Das Gewicht für den prozentualen GC-Gehalt wird wie folgt berechnet:

$$GCScore = |\langle GC \rangle - GC_{Wunsch}|^{1.3} \times 0,8$$

Zur Ermittlung der Einzelgewichte der Alignments (Alignmentscore) wird ein optimales lokales Alignment der Testsequenz mit einem Teilbereich der Testsequenz, der maximal die letzten 36 Basen der kompletten Testsequenz umfasst, unter Ausschluss des Identitätsalignments (Alignment des vollständigen Teilbereiches mit sich selbst) generiert (vgl. Fig. 3, 4a, 4b).

Als Bewertungsparameter für eine Basenposition zur Berechnung der Dynamic-Programming Matrix werden dabei verwendet:

Übereinstimmung (Match) = 1;

Fehlpaarung (Mismatch) = -2;

Lücke (Gap) = -2.

Das entsprechende Kriteriumsgewicht wird durch eine Potenz des optimalen Alignment-Scores in dem überprüften Bereich der Testsequenz festgelegt:

$$REPScore = (Score_{Alignment})^{1.3}$$

Für jede gefundene CATG-Sequenz wird ein Sitescore von 100000 vergeben.

Die Gesamtgütefunktion GesScore ergibt sich

$$GesScore = CUScore - GCscore - REPScore - SiteScore$$

Die KDS-Länge m beträgt 3 Codons (9 Basen).

Eine Optimierung lediglich auf optimale Codon-Usage resultiert in folgender Sequenz:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
E	Q	F	I	I	K	N	M	F	I	I	K	N	A
GAA	CAG	TTT	ATT	ATT	AAA	AAC	ATG	TTT	ATT	ATT	AAA	AAC	GCG

Sie ist durch folgende Eigenschaften charakterisiert:

Stark repetitiv, verursacht durch die zweimalig erscheinende Aminosäuresequenz F_I_I_K_N (gezeigt ist das repetitive Element mit dem höchsten Score (18)):

```

19 AACATGTTTATTATTAAAAAC
   |||| |
2  AACACGTTTATTATTAAAAAC

```

- GC-Gehalt: 21,4 %
- Die Nla III Erkennungssequenz CATG ist vorhanden
- Durchschnittliche Codon-Usage: 100

Wird die Optimierung nach dem erfindungsgemäßen Algorithmus mit den oben genannten Bewertungsfunktionen und Parametern vorgenommen, so erhält man folgende DNS-Sequenz:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
E__	Q__	F__	I__	I__	K__	N__	M__	F__	I__	I__	K__	N__	A__
GAA	CAG	TTC	ATC	ATC	AAA	AAT	ATG	TTT	ATT	ATC	AAG	AAC	GCG

Sie ist durch folgende Eigenschaften charakterisiert:

- Kaum repetitiv (das nachfolgend gezeigte Alignment mit dem höchsten Beitrag hat einen Score 6)

```

11 TCATCA
   |||||
 8 TCATCA

```

- GC-Gehalt: 31,0 %
- Die Nla III Erkennungssequenz CATG ist vermieden worden
- Durchschnittliche Codon-Usage: 88

Bei dem erfindungsgemäßen Optimierungsergebnis wurde an fünf Aminosäure-Positionen nicht das hinsichtlich der Codon usage optimale Codon gewählt. Die erfindungsgemäße aufgefundene Sequenz stellt jedoch eine optimale Balance der unterschiedlichen Anforderungen in Bezug auf Codon-Usage, GC-Gehalt und ideale Sequenzeigenschaften (Vermeidung von Repetitionen) dar.

Bei den Aminosäuren mit den Nummern 3,4,5 ist der höhere GC-Anteil der hinsichtlich der Codon usage schlechteren Codons der Grund für die Wahl. An Position 6 überwiegt jedoch beim Vergleich der Codons AAA und AAG die wesentlich bessere Codon usage des AAA Codons, obwohl die Wahl des AAG Codons zu einem besseren GC-Score führen würde. Bei Bildung der KDS an Basenposition 13 wird für die Aminosäure Nr. 7 noch das Codon AAC

bevorzugt, da bei einer Fenstergröße für die KDS von 3 Codons noch nicht erkennbar ist, daß diese Wahl zu Bildung des zu vermeidenden DNS-Motivs CATG führen wird (für Methionin ist der genetische Code nicht degeneriert, d.h. es gibt nur ein Codon zur Expression von Methionin). Bei der Bildung der KDS an Basenposition 16 wird dies jedoch bereits erkannt und folgerichtig das Codon AAT gewählt. Bei der Wahl der Codons für die Aminosäuren 9 bis 13 spielt neben Codon-Usage und GC-Gehalt auch die Vermeidung einer repetitiven DNS-Sequenz. Aufgrund der identischen Aminosäuresequenzen der Aminosäuren Nr. 3 bis 7 und 9 bis 13 eine entscheidende Rolle. Aus diesem Grunde werden für die Aminosäuren 9 und 10 im Gegensatz zu vorher (Asr. 3,4) die Codons TTT und ATT bevorzugt.

Die am Ende der Beschreibung beigefügte Tabelle illustriert die einzelnen Schritte des Algorithmus, die zu dem oben angegebenen Optimierungsergebnis geführt haben. Sie ermöglicht es, den Ablauf des Algorithmus Schritt für Schritt nachzuvollziehen. Für jede Startposition werden dabei detailliert alle von der Software gebildeten Kombinations-DNS-Sequenzen (KDS) aufgelistet.

Zu jeder möglichen KDS werden folgende Angaben gemacht:

- die aus der jeweiligen KDS und der bereits optimierten DNS-Sequenz gebildete Testsequenz, welche zur Evaluierung der KDS herangezogen wird,
- die Scores, welche für Codon usage, GC-Gehalt, Repetitivität und aufgefundenen DNS-Sites ermittelt wurden (CU, GC, Rep, Site)
- das für die jeweilige Testsequenz ermittelte repetitive Element mit dem höchsten Alignment-Score,
- der ermittelte Gesamtscore.

Die KDS sind dabei nach fallendem Gesamtscore sortiert, d.h. das erste Codon der ersten gezeigten KDS wird an die bereits optimierte DNS-Sequenz angefügt.

Beispiel 2

Bei diesem Beispiel wird die Optimierung von GFP auf Expression in E. Coli betrachtet.

Herkunft der Aminosäuresequenz:

DEFINITION Aequorea victoria green-fluorescent protein mRNA, complete cds.
 ACCESSION M62654

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTCLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYP
 DHMKQHDFFKSAMPEGYVQERTIFYKDDGNYKSRAEVKFEGLTLVNRIELKGIDFKEDGNILGHKMEYNYNSHNV
 YIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMILLEFVT
 AAGITHGMDELYK

Verwendete Codon-Usage-Table: Escherichia coli K12

Herkunft : Codon usage Database auf www.kazusa.or.jp/codon

Nachfolgend bedeuten:

<CU> : durchschnittliche renormierte Codon-Usage der KDS (15 Basen lang)

<GC> : durchschnittlicher prozentualer GC-Gehalt der letzten 35 Basen der Testsequenz

GC_{Wunsch}: Angestrebter GC-Gehalt

Die Größe des Fenster, auf dem der GC-Gehalt für die graphische Darstellung in Fig. 5b bis 8b berechnet wurde, betrug 40 Basen

Fig. 5a und 5b zeigen die Ergebnisse für die Gütefunktion:

$$Score = \langle CU \rangle$$

Fig. 6a und 6b zeigen die Ergebnisse für die Gütefunktion

$$Score = \langle CU \rangle - |\langle GC \rangle - GC_{Wunsch}|^{1.3} \times 0.8$$

Fig. 7a und 7b zeigen die Ergebnisse für die Gütefunktion

$$Score = \langle CU \rangle - |\langle GC \rangle - GC_{Wunsch}|^{1.3} \times 1.5$$

Fig. 8a und 8b zeigen die Ergebnisse für die Gütefunktion

$$Score = \langle CU \rangle - |\langle GC \rangle - GC_{Wunsch}|^{1.3} \times 5$$

Die Figuren 5 bis 8 verdeutlichen den Einfluß der unterschiedlichen Gewichtung zweier Optimierungskriterien auf das Optimierungsergebnis. Ziel ist, die GC-Gehaltsverteilung über die Sequenz zu glätten und den Wert 50% anzunähern. In dem in Fig. 5a und 5b gezeigten Fall wurde lediglich auf die optimale Codon-Usage optimiert, was in einer sehr heterogenen und vom Ziel-Gehalt teilweise stark abweichenden GC-Verteilung resultiert. In dem Fall der Fig. 6a und 6b verbindet sich in idealer Weise eine Glättung des GC-Gehaltes auf einen Wert um 50% mit einer guten bis sehr guten Codon-Usage. Die Fälle der Fig. 7a und 7b bzw. 8a und 8b verdeutlichen schließlich, daß eine weitere GC-Gehalts-Optimierung zwar möglich ist, aber mit einer stellenweise schlechten Codon-Usage erkauft werden muß.

Die in den Ansprüchen, den Zeichnungen und der Beschreibung offenbarten Merkmale können sowohl einzeln als auch in beliebiger Kombination für die Verwirklichung der Erfindung in ihren verschiedenen Ausführungsformen wesentlich sein.

KDS-Startposition 1 bei Aminosäure 1 E

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
GAACAGTTC GAACAGTTC	92	5	0	0,0	G G	87,0
GAACAGTTT GAACAGTTT	100	19	0	0,0	TT TT	81,0
GAGCAGTTT GAGCAGTTT	82	5	0	0,0	AG AG	77,0
GAGCAGTTC GAGCAGTTC	73	5	0	0,0	AG AG	68,0
GAACAATTC GAACAATTC	76	19	0	0,0	AA AA	57,0
GAGCAATTC GAGCAATTC	58	5	0	0,0	G G	53,0
GAACAATTT GAACAATTT	85	38	0	0,0	AA AA	47,0
GAGCAATTT GAGCAATTT	66	19	0	0,0	TT TT	47,0

KDS-Startposition 4 bei Aminosäure 2 Q

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
CAGTTCATC GAACAGTTCATC	86	8	0	0,0	CA CA	78,0
CAGTTTATC GAACAGTTTATC	94	19	0	0,0	TT TT	75,0
CAGTTCATT GAACAGTTCATT	92	19	0	0,0	CA CA	73,0
CAGTTTATT GAACAGTTTATT	100	33	0	0,0	TT TT	67,0
CAATTCATC GAACAATTCATC	70	19	0	0,0	AA AA	51,0
CAATTTATC GAACAATTTATC	79	33	0	0,0	AA AA	46,0
CAGTTCATA GAACAGTTCATA	63	19	0	0,0	CA CA	44,0
CAATTCATT GAACAATTCATT	76	33	0	0,0	ATT ATT	43,0
CAGTTTATA GAACAGTTTATA	71	33	0	0,0	TT TT	38,0
CAATTTATT GAACAATTTATT	85	48	0	0,0	ATT ATT	37,0
CAATTCATA GAACAATTCATA	48	33	0	0,0	AA AA	15,0
CAATTTATA GAACAATTTATA	56	48	0	0,0	AA AA	8,0

KDS-Startposition 7 bei Aminosäure 3 F

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
TCATCATC GAACAGTTCATCATC	80	10	0	0,0	TCATC TCATC	70,0

TTTATCATC 88 GAACAGTTTATCATC	19	0	0,0	ATC ATC	69,0
TTCATTATC 86 GAACAGTTCATTATC	19	0	0,0	CA CA	67,0
TTCATCATT 86 GAACAGTTCATCATT	19	0	0,0	TCAT TCAT	67,0
TTTATTATC 94 GAACAGTTTATTATC	30	0	0,0	TTAT TTAT	64,0
TTTATCATT 94 GAACAGTTTATCATT	30	0	0,0	CA CA	64,0
TTCATTATT 92 GAACAGTTCATTATT	30	0	0,0	ATT ATT	62,0
TTTATTATT 100 GAACAGTTTATTATT	42	0	0,0	TTATT TTATT	58,0
TTCATCATA 57 GAACAGTTCATCATA	19	0	0,0	TCAT TCAT	38,0
TTCATAATC 57 GAACAGTTCATAATC	19	0	0,0	AA AA	38,0
TTTATCATA 65 GAACAGTTTATCATA	30	0	0,0	CA CA	35,0
TTATAATC 65 GAACAGTTTATAATC	30	0	0,0	AA AA	35,0
TTCATTATA 63 GAACAGTTCATTATA	30	0	0,0	CA CA	33,0
TTCATAATT 63 GAACAGTTCATAATT	30	0	0,0	AA AA	33,0
TTTATTATA 71 GAACAGTTTATTATA	42	0	0,0	TTAT TTAT	29,0
TTTATAATT 71 GAACAGTTTATAATT	42	0	0,0	AA AA	29,0
TTCATAATA 34 GAACAGTTCATAATA	30	0	0,0	ATA ATA	4,0
TTTATAATA 43 GAACAGTTTATAATA	42	0	0,0	ATA ATA	1,0

KDS-Startposition 10 bei Aminosäure 4 I

PS estsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
ATCATCAAA GAACAGTTCATCATCAAA	88	19	0	0,0	TCATCA TCATCA	69,0
ATTATCAAA GAACAGTTCATTATCAAA	94	28	0	0,0	TCA TCA	66,0
ATCATTAAG GAACAGTTCATCATTAAG	94	28	0	0,0	TCAT TCAT	66,0
ATTATTAAG GAACAGTTCATTATTAAG	100	38	0	0,0	ATTA ATTA	62,0
ATCATCAAG GAACAGTTCATCATCAAG	65	11	0	0,0	TCATCA TCATCA	54,0
ATTATCAAG GAACAGTTCATTATCAAG	71	19	0	0,0	TCA TCA	52,0
ATCATTAAG GAACAGTTCATCATTAAG	71	19	0	0,0	TCAT TCAT	52,0
ATTATTAAG GAACAGTTCATTATTAAG	77	28	0	0,0	ATTA ATTA	49,0

ATCATAAAA 65	28	0	0,0	TCAT TCAT	37,0
GAACAGTTCATCATAAAA					
ATAATCAAAA 65	28	0	0,0	TCA TCA	37,0
GAACAGTTCATAATCAAA					
ATTATAAAA 71	38	0	0,0	AAA AAA	33,0
GAACAGTTCATTATAAAA					
ATAATTA AAA 71	38	0	0,0	TAA TAA	33,0
GAACAGTTCATAATTA AA					
ATCATAAAG 43	19	0	0,0	TCAT TCAT	24,0
GAACAGTTCATCATAAAG					
ATAATCAAG 43	19	0	0,0	TCA TCA	24,0
GAACAGTTCATAATCAAG					
ATTATAAAG 49	28	0	0,0	AA AA	21,0
GAACAGTTCATTATAAAG					
ATAATTAAG 49	28	0	0,0	TAA TAA	21,0
GAACAGTTCATAATTAAG					
ATAATAAAA 43	38	0	0,0	ATAA ATAA	5,0
GAACAGTTCATAATAAAA					
ATAATAAAG 20	28	0	0,0	ATAA ATAA	-8,0
GAACAGTTCATAATAAAG					

KDS-Startposition 13 bei Aminosäure 5 I

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
ATCAAAAAC GAACAGTTCATCATCAAAAAC	94	19	0	0,0	TCATCA TCATCA	75,0
ATTA AAAAC GAACAGTTCATCATTAAAAAC	100	27	0	0,0	TCAT TCAT	73,0
ATCAAAAAT GAACAGTTCATCATCAAAAAT	88	27	0	0,0	TCATCA TCATCA	61,0
ATTA AAAAT GAACAGTTCATCATTAAAAAT	94	35	0	0,0	TCAT TCAT	59,0
ATTAAGAAC GAACAGTTCATCATTAGAAC	77	19	0	0,0	GAAC GAAC	58,0
ATCAAGAAC GAACAGTTCATCATCAAGAAC	71	13	0	0,0	TCATCA TCATCA	58,0
CAAGAAT GAACAGTTCATCATCAAGAAT	65	19	0	0,0	TCATCA TCATCA	46,0
ATTAAGAAT GAACAGTTCATCATTAGAAT	71	27	0	0,0	TCAT TCAT	44,0
ATAAAAAAC GAACAGTTCATCATAAAAAAC	71	27	0	0,0	TCAT-A-A-A-A-A TCATCATAAAAA	44,0
ATAAAAAAT GAACAGTTCATCATAAAAAAT	65	35	0	0,0	TCAT-A-A-A-A-A TCATCATAAAAA	30,0
ATAAAGAAC GAACAGTTCATCATAAAGAAC	49	19	0	0,0	GAAC GAAC	30,0
ATAAAGAAT GAACAGTTCATCATAAAGAAT	43	27	0	0,0	TCAT TCAT	16,0

KDS-Startposition 16 bei Aminosäure 6 K

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
AAAAATATG	94	26	0	0,0	TCATCA	68,0
GAACAGTTCATCATCAAAAATATG					TCATCA	

AAGAATATG	71	19	0	0,0	TCATCA	52,0
GAACAGTTCATCATCAAGAATATG					TCATCA	
AAAAACATG	100	19	200000	0,0	TCATCA	919,0
GAACAGTTCATCATCAAAAACATG					TCATCA	
AAGAACATG	77	13	200000	0,0	TCATCA	936,0
GAACAGTTCATCATCAAGAACATG					TCATCA	

KDS-Startposition 19 bei Aminosäure 7 N

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
AATATGTTT	94	35	0	0,0	TCATCA	59,0
GAACAGTTCATCATCAAAAATATGTTT					TCATCA	
AATATGTTC	86	28	0	0,0	TCATCA	58,0
GAACAGTTCATCATCAAAAATATGTTC					TCATCA	
AACATGTTT	100	28	200000	0,0	TCATCA	928,0
GAACAGTTCATCATCAAAAACATGTTT					TCATCA	
AACATGTTC	92	21	200000	0,0	AACATGTTT	929,0
GAACAGTTCATCATCAAAAACATGTTC					AACA-GTTC	

KDS-Startposition 22 bei Aminosäure 8 M

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
ATGTTTATC	94	35	0	0,0	TCATCA	59,0
GAACAGTTCATCATCAAAAATATGTTTATC					TCATCA	
ATGTTTATT	100	42	0	0,0	TCATCA	58,0
GAACAGTTCATCATCAAAAATATGTTTATT					TCATCA	
ATGTTTCATT	92	35	0	0,0	GTTTCAT	57,0
GAACAGTTCATCATCAAAAATATGTTTCATT					GTTTCAT	
ATGTTTCATC	86	28	0	12,5	GTTTCATC	45,0
GAACAGTTCATCATCAAAAATATGTTTCATC					GTTTCATC	
ATGTTTATA	71	42	0	0,0	TCATCA	29,0
GAACAGTTCATCATCAAAAATATGTTTATA					TCATCA	
ATGTTTCATA	63	35	0	0,0	GTTTCAT	28,0
GAACAGTTCATCATCAAAAATATGTTTCATA					GTTTCAT	

KDS-Startposition 25 bei Aminosäure 9 F

KDS Testsequenz	CU	GC	Site	Rep	Alignment	Gesamt-Score
TTTATTATC	94	42	0	0,0	TCATCA	52,0
GAACAGTTCATCATCAAAAATATGTTTATTATC					TCATCA	
TTTATCATT	94	42	0	0,0	TCATCA	52,0
GAACAGTTCATCATCAAAAATATGTTTATCATT					TCATCA	
TTTATTATT	92	42	0	0,0	GTTTCAT	50,0
GAACAGTTCATCATCAAAAATATGTTTCATTATT					GTTTCAT	
TTTATCATC	88	35	0	12,5	GTTTATCATC	40,0
GAACAGTTCATCATCAAAAATATGTTTATCATC					GTTTATCATC	
TTTATTATT	100	49	0	12,5	TCATCA--AAAATATGTTTATTATT	38,0
GAACAGTTCATCATCAAAAATATGTTTATTATT					TCATCATCAAAA-ATATGT-TTATT	
TTTATTATC	86	35	0	12,5	GTTTATTATC	38,0
GAACAGTTCATCATCAAAAATATGTTTCATTATC					GTTTATTATC	
TTTATCATT	86	35	0	17,4	GTTTATCAT	34,0
GAACAGTTCATCATCAAAAATATGTTTCATCATT					GTTTATCAT	

TTCATCATC	80	28	0	20,0	GTTTCATCATC	32,0
GAACAGTTCATCATCAAAAATATGTTTCATCATC					GTTTCATCATC	
TTTATCATA	65	42	0	0,0	TCATCA	23,0
GAACAGTTCATCATCAAAAATATGTTTATCATA					TCATCA	
TTTATAATC	65	42	0	0,0	TCATCA	23,0
GAACAGTTCATCATCAAAAATATGTTTATAATC					TCATCA	
TTTATTATA	71	49	0	0,0	TCATCA	22,0
GAACAGTTCATCATCAAAAATATGTTTATTATA					TCATCA	
TTTATAATT	71	49	0	0,0	TCATCA	22,0
GAACAGTTCATCATCAAAAATATGTTTATAATT					TCATCA	
TTCATAATT	63	42	0	0,0	GTTTCAT	21,0
GAACAGTTCATCATCAAAAATATGTTTCATAATT					GTTTCAT	
TTCATTATA	63	42	0	0,0	GTTTCAT	21,0
GAACAGTTCATCATCAAAAATATGTTTCATTATA					GTTTCAT	
TTCATAATC	57	35	0	12,5	GTTTCATAATC	9,0
GAACAGTTCATCATCAAAAATATGTTTCATAATC					GTTTCATCATC	
TTCATCATA	57	35	0	17,4	GTTTCATCAT	5,0
GAACAGTTCATCATCAAAAATATGTTTCATCATA					GTTTCATCAT	
TTTATAATA	43	49	0	0,0	TCATCA	-6,0
GAACAGTTCATCATCAAAAATATGTTTATAATA					TCATCA	
TCATAATA	34	42	0	0,0	GTTTCAT	-8,0
GAACAGTTCATCATCAAAAATATGTTTCATAATA					GTTTCAT	

KDS-Startposition 28 bei Aminosäure 10 I

KDS	CU	GC	Site	Rep	Alignment	Gesamt-Score
Testsequenz						
ATTATCAAA	94	49	0	12,5	GTTTATTATCAAA	32,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAA					GTTTCATCATCAAA	
ATCATTAATA	94	49	0	12,5	GTTTATCATTAATA	32,0
GAACAGTTCATCATCAAAAATATGTTTATCATTAA					GTTTCATCATCAAA	
ATTATCAAG	71	42	0	0,0	TCATCA	29,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAG					TCATCA	
ATCATTAAG	71	42	0	0,0	TCATCA	29,0
GAACAGTTCATCATCAAAAATATGTTTATCATTAA					TCATCA	
ATTATTAATA	100	57	0	14,9	TCATCA--AAAATATGTTTATTATTA	28,0
GAACAGTTCATCATCAAAAATATGTTTATTATTAATA					TCATCATCAAAA-ATATGT-TTATTA	
TCATCAAA	88	42	0	20,0	GTTTATCATCAAA	26,0
GAACAGTTCATCATCAAAAATATGTTTATCATCAAA					GTTTCATCATCAAA	
ATTATAAAA	71	57	0	0,0	TCATCA	14,0
GAACAGTTCATCATCAAAAATATGTTTATTATAAAA					TCATCA	
ATAATTAATA	71	57	0	0,0	TCATCA	14,0
GAACAGTTCATCATCAAAAATATGTTTATAATTAA					TCATCA	
ATTATTAAG	77	49	0	14,9	TCATCA--AAAATATGTTTATTATTA	13,0
GAACAGTTCATCATCAAAAATATGTTTATTATTAAG					TCATCATCAAAA-ATATGT-TTATTA	
ATCATCAAG	65	35	0	17,4	GTTTATCATCAA	13,0
GAACAGTTCATCATCAAAAATATGTTTATCATCAAG					GTTTCATCATCAA	
ATAATCAAA	65	49	0	12,5	GTTTATAATCAAA	3,0
GAACAGTTCATCATCAAAAATATGTTTATAATCAAA					GTTTCATCATCAAA	
ATCATAAAA	65	49	0	14,9	GTTTATCAT-AAAA	1,0
GAACAGTTCATCATCAAAAATATGTTTATCATAAAA					GTTTCATCATCAAAA	
ATAATCAAG	43	42	0	0,0	TCATCA	1,0
GAACAGTTCATCATCAAAAATATGTTTATAATCAAG					TCATCA	
ATTATAAAG	49	49	0	0,0	TCATCA	0,0
GAACAGTTCATCATCAAAAATATGTTTATTATAAAG					TCATCA	

ATAATTAAG	49	49	0	0,0	TCATCA	0,0
GAACAGTTCATCATCAAAAATATGTTTATAATTAAG					TCATCA	
ATCATAAAG	43	42	0	12,5	GTTTATCAT-AAA	-12,0
GAACAGTTCATCATCAAAAATATGTTTATCATAAAG					GTTTATCATCAAAA	
ATAATAAAA	43	57	0	0,0	TCATCA	-14,0
GAACAGTTCATCATCAAAAATATGTTTATAATAAAA					TCATCA	
ATAATAAAG	20	49	0	0,0	TCATCA	-29,0
GAACAGTTCATCATCAAAAATATGTTTATAATAAAG					TCATCA	

KDS-Startposition 31 bei Aminosäure 11 I

KDS	CU	GC	Site	Rep	Alignment	Gesamt-Score
Testsequenz						
ATCAAGAAC	71	42	0	0,0	TCATCA	29,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAAC					TCATCA	
ATTA AAAAC	100	57	0	14,9	TCATCA--AAAATATGTTTATTATTA	28,0
GAACAGTTCATCATCAAAAATATGTTTATTATTA AAAAC					TCATCATCAAAA-ATATGT-TTATTA	
ATCA AAAAC	94	49	0	17,4	GTTTATTATCA AAAA	28,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAAC					GTTTATCATCA AAAA	
TTA AAAAT	94	64	0	14,9	TCATCA--AAAATATGTTTATTATTA	15,0
GAACAGTTCATCATCAAAAATATGTTTATTATTA AAAAT					TCATCATCAAAA-ATATGT-TTATTA	
ATTAAGAAC	77	49	0	14,9	TCATCA--AAAATATGTTTATTATTA	13,0
GAACAGTTCATCATCAAAAATATGTTTATTATTAAGAAC					TCATCATCAAAA-ATATGT-TTATTA	
ATCA AAAAT	88	57	0	20,0	GTTTATTATCA AAAAT	11,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAAT					GTTTATCATCA AAAAT	
ATCAAGAAT	65	49	0	12,5	GTTTATTATCAAGAAT	3,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAAT					GTTTATCATCA AAAAT	
ATAAAGAAC	49	49	0	0,0	TCATCA	0,0
GAACAGTTCATCATCAAAAATATGTTTATTATAAAGAAC					TCATCA	
ATTAAGAAT	71	57	0	14,9	TCATCA--AAAATATGTTTATTATTA	-1,0
GAACAGTTCATCATCAAAAATATGTTTATTATTAAGAAT					TCATCATCAAAA-ATATGT-TTATTA	
ATAA AAAAC	71	57	0	14,9	TCATCA--AAAATATGTTTATTA-TA-AAAAA	-1,0
GAACAGTTCATCATCAAAAATATGTTTATTATAA AAAAC					TCATCATCAAAA-ATATGT-TTATTATAAAAAA	
ATAA AAAAT	65	64	0	14,9	TCATCA--AAAATATGTTTATTA-TA-AAAAA	-14,0
GAACAGTTCATCATCAAAAATATGTTTATTATAA AAAAT					TCATCATCAAAA-ATATGT-TTATTATAAAAAA	
ATAAAGAAT	43	57	0	0,0	TCATCA	-14,0
GAACAGTTCATCATCAAAAATATGTTTATTATAAAGAAT					TCATCA	

KDS-Startposition 34 bei Aminosäure 12 K

KDS	CU	GC	Site	Rep	Alignment	Gesamt-Score
Testsequenz						
AAGAACGCG	77	28	0	0,0	TCATCA	49,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAACGCG					TCATCA	
AAAAACGCG	100	35	0	17,4	GTTTATTATCA AAAA	48,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAACGCG					GTTTATCATCA AAAA	
AAGAACGCC	69	28	0	0,0	TCATCA	41,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAACGCC					TCATCA	
AAAAACGCC	92	35	0	17,4	GTTTATTATCA AAAA	40,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAACGCC					GTTTATCATCA AAAA	
AAAAATGCG	94	42	0	20,0	GTTTATTATCA AAAAT	32,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAATGCG					GTTTATCATCA AAAAT	
AAGAACGCA	63	35	0	0,0	TCATCA	28,0
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAACGCA					TCATCA	
AAAAACGCA	86	42	0	17,4	GTTTATTATCA AAAA	27,0
GAACAGTTCATCATCAAAAATATGTTTATTATCA AAAACGCA					GTTTATCATCA AAAA	

AAAAATGCC	86	42	0	20,0	GTTTATTATCAAAAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAAAATGCC					GTTTCATCATCAAAAAT	24,0
AAGAACGCT	59	35	0	0,0	TCATCA	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAACGCT					TCATCA	24,0
AAGAATGCG	71	35	0	12,5	GTTTATTATCAAGAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAATGCG					GTTTCATCATCAAAAAT	23,0
AAAAACGCT	81	42	0	17,4	GTTTATTATCAAAAA	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAAAACGCT					GTTTCATCATCAAAAA	22,0
AAGAATGCC	63	35	0	12,5	GTTTATTATCAAGAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAATGCC					GTTTCATCATCAAAAAT	15,0
AAAAATGCA	80	49	0	20,0	GTTTATTATCAAAAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAAAATGCA					GTTTCATCATCAAAAAT	11,0
AAAAATGCT	75	49	0	20,0	GTTTATTATCAAAAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAAAATGCT					GTTTCATCATCAAAAAT	6,0
AAGAATGCA	57	42	0	12,5	GTTTATTATCAAGAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAATGCA					GTTTCATCATCAAAAAT	2,0
AAGAATGCT	53	42	0	12,5	GTTTATTATCAAGAAT	
GAACAGTTCATCATCAAAAATATGTTTATTATCAAGAATGCT					GTTTCATCATCAAAAAT	-2,0

Zusammenfassung

Die Erfindung betrifft ein Verfahren zum Optimieren einer Nucleotidsequenz zur Expression eines Proteins auf der Grundlage der Aminosäuresequenz des Proteins, bei welchem für einen bestimmten Bereich eine Testsequenz mit m Optimierungspositionen festgelegt wird, auf denen die Codonbesetzung variiert wird, wobei mittels einer Gütefunktion die optimale Codonbesetzung auf diesen Optimierungspositionen ermittelt wird und ein oder mehrere Codons dieser optimalen Besetzung als Codons der optimierten Nucleotidsequenz festgelegt werden. Diese Schritte werden iteriert, wobei bei nachfolgenden Iterationsschritten die in vorangehenden Schritten festgelegten Codons der optimierten Nucleotidsequenz unverändert bleiben. Die Erfindung betrifft weiterhin eine Vorrichtung zur Durchführung dieses Verfahrens.

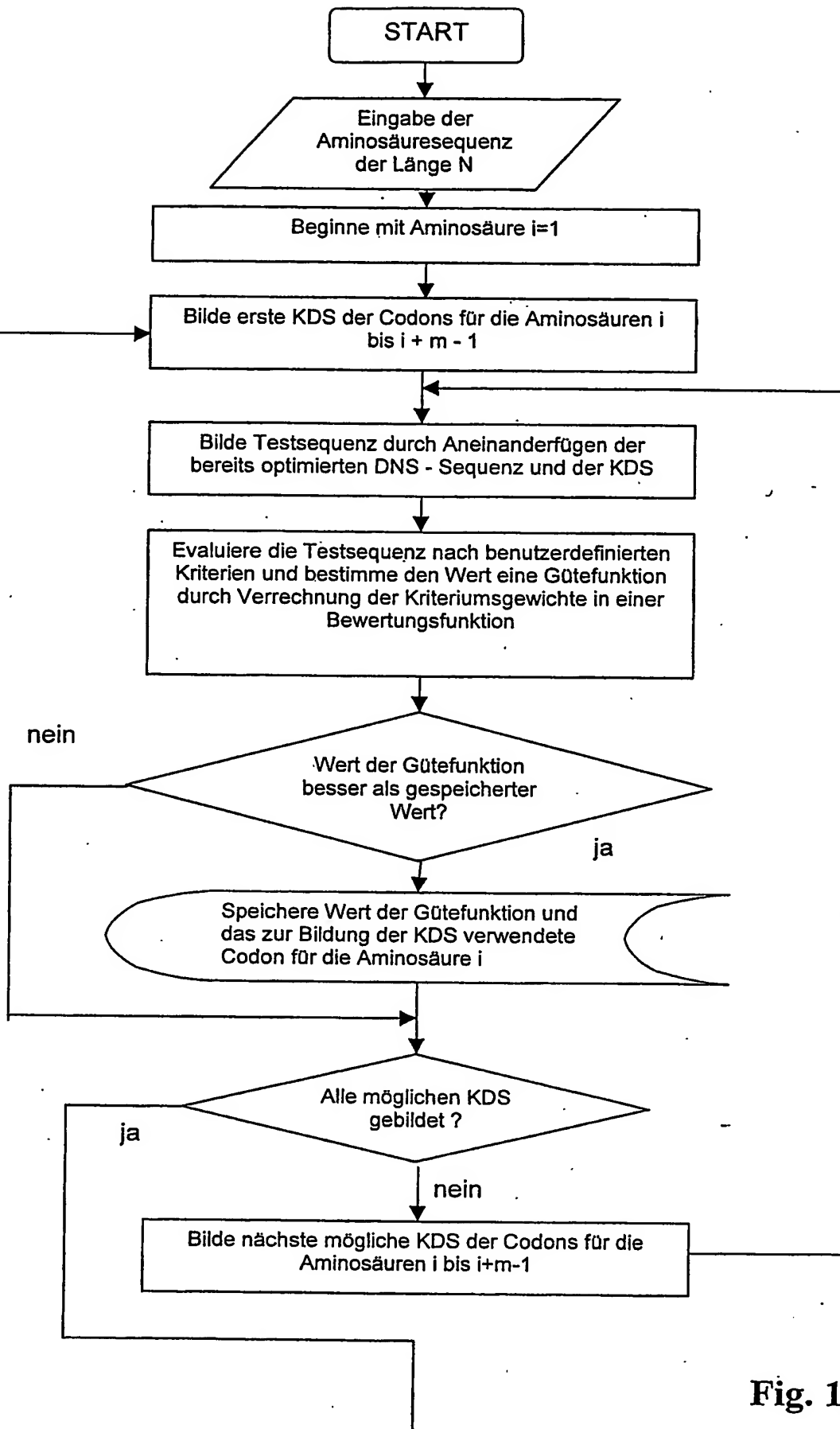


Fig. 1a

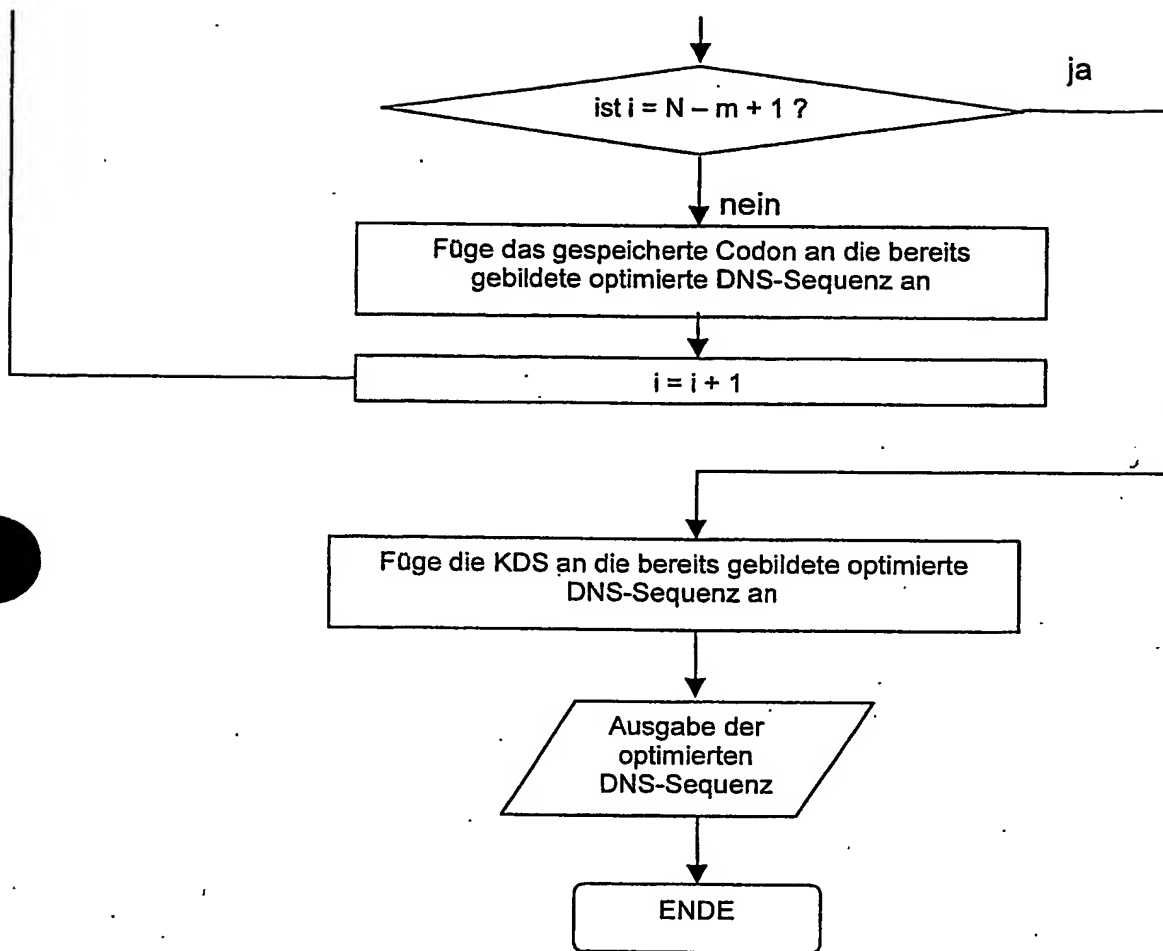


Fig. 1b

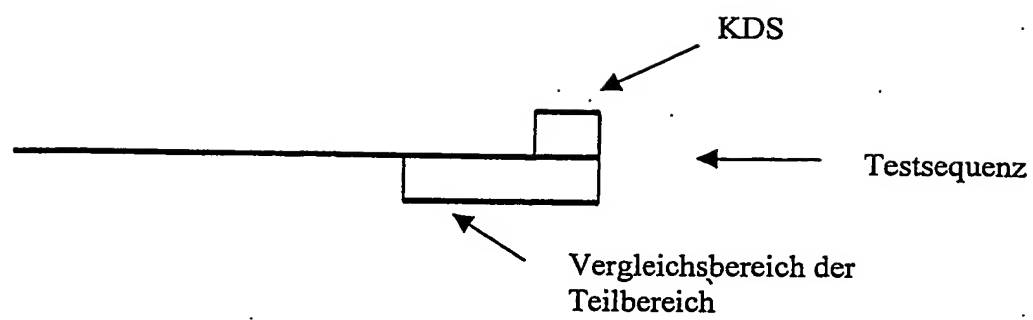


Fig. 3

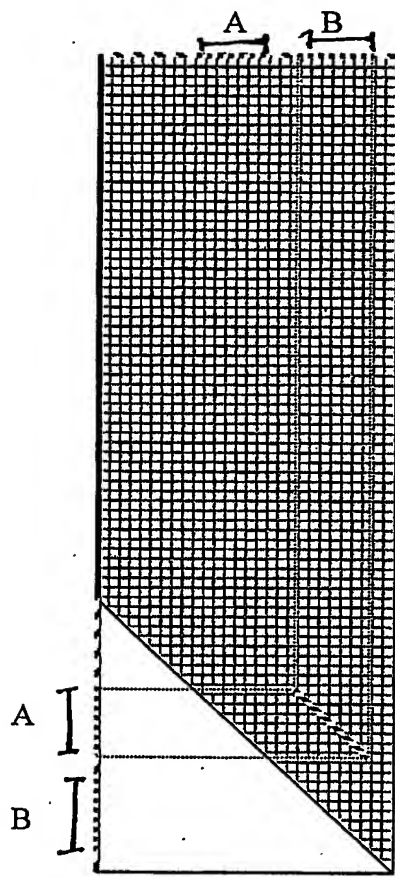


Fig. 4a

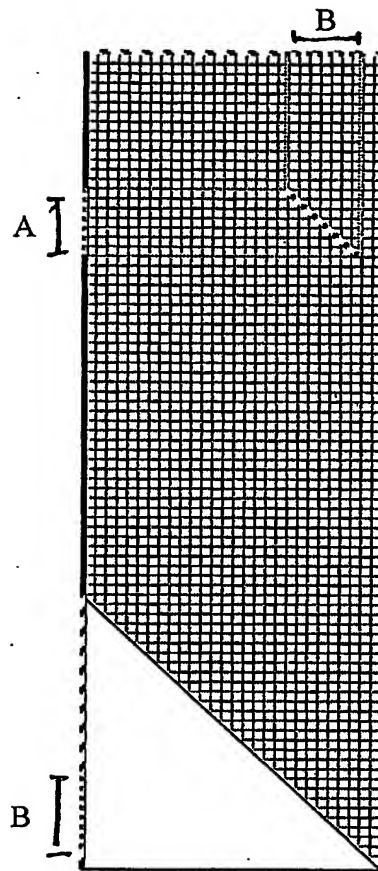


Fig. 4b

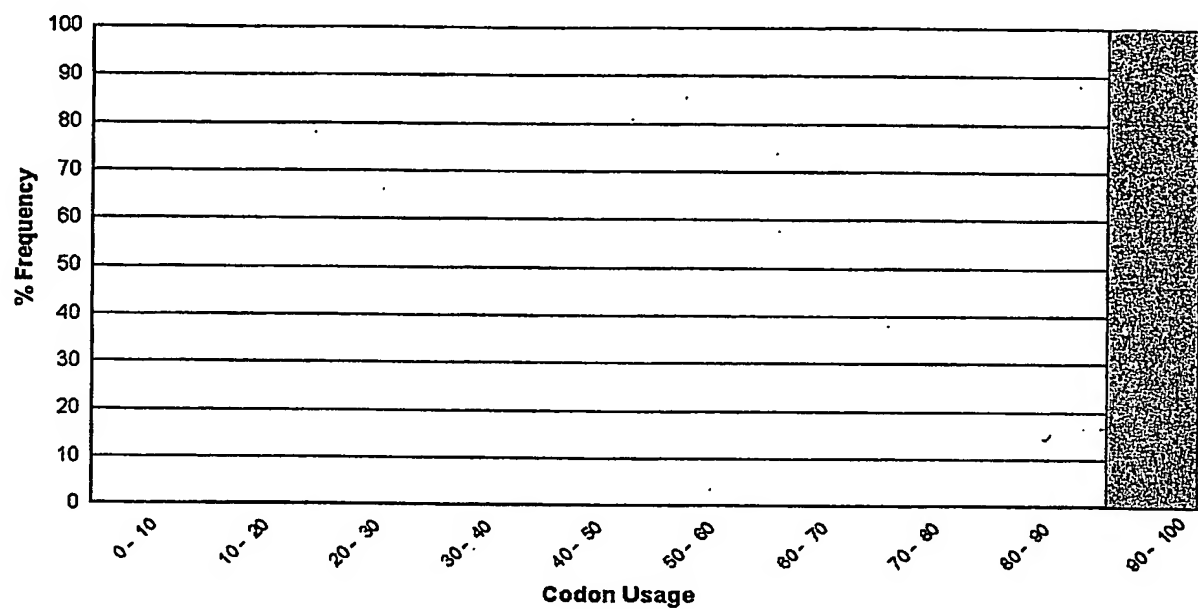


Fig. 5a

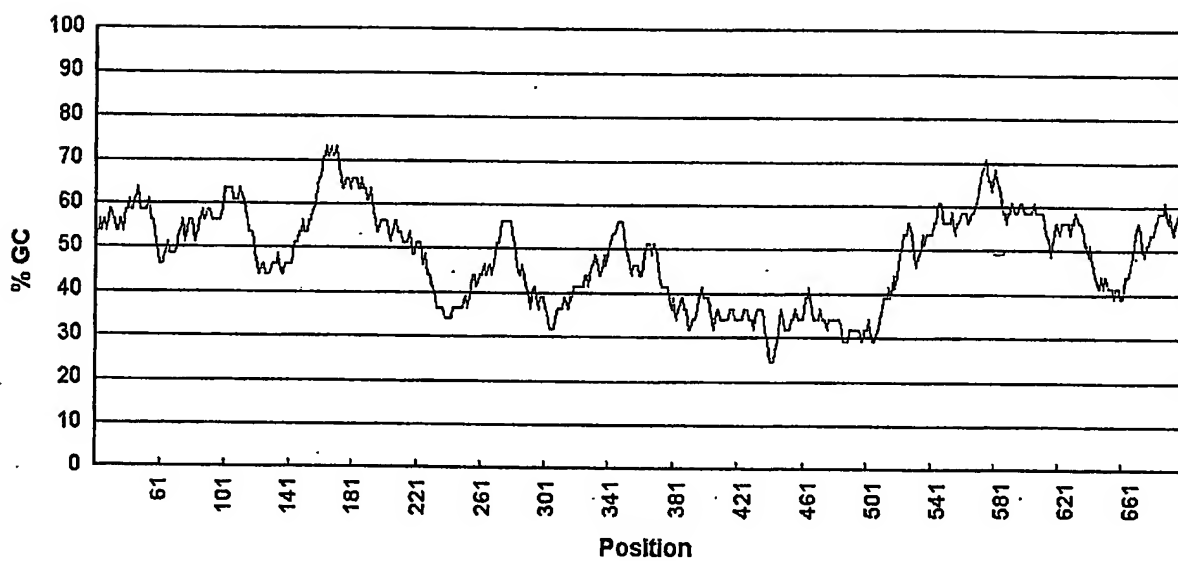


Fig. 5b

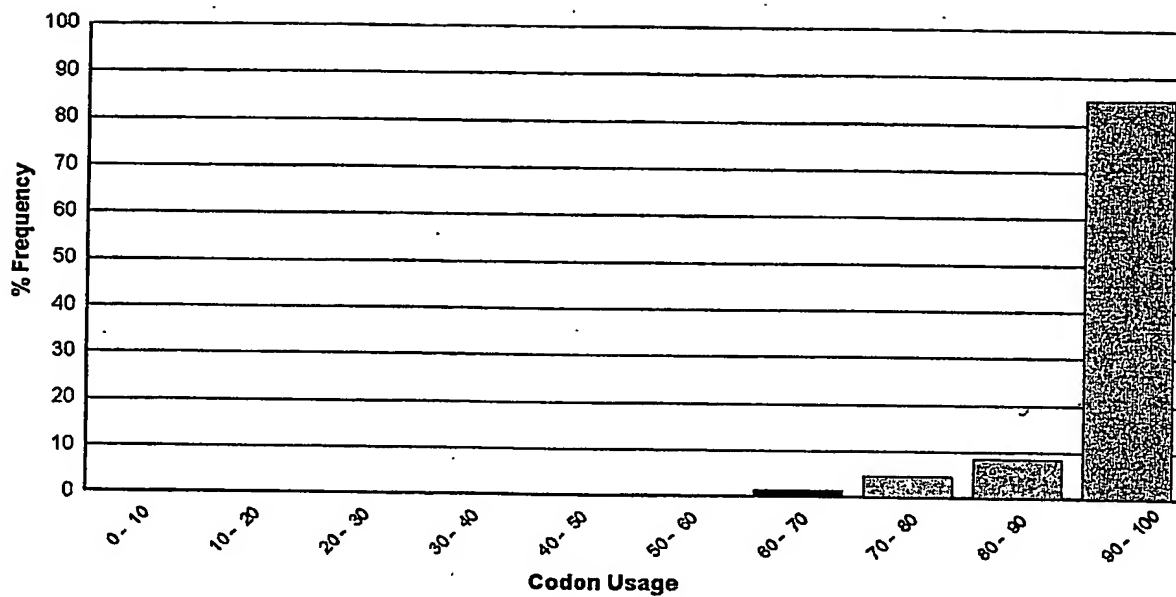


Fig. 6a

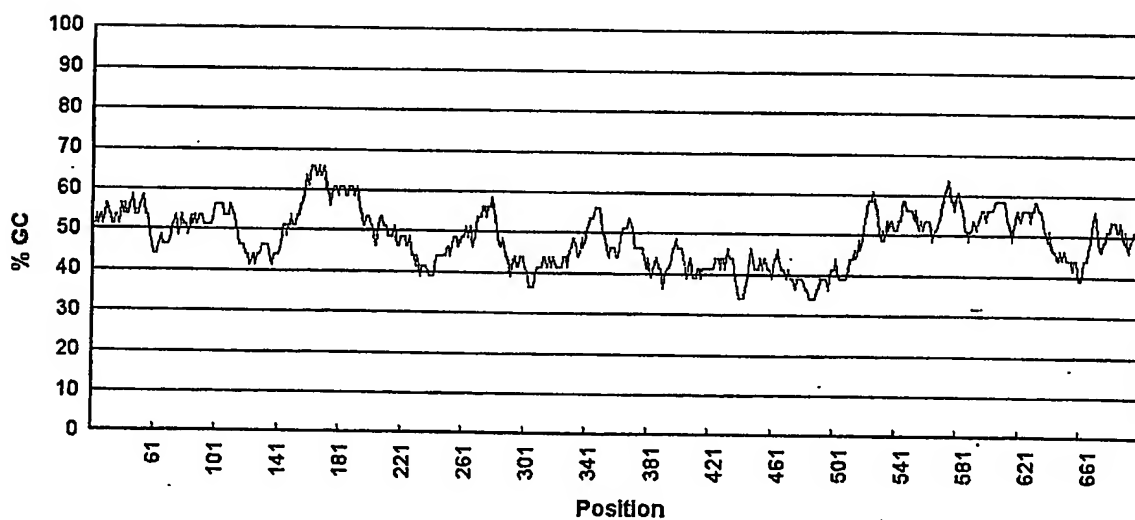


Fig. 6b

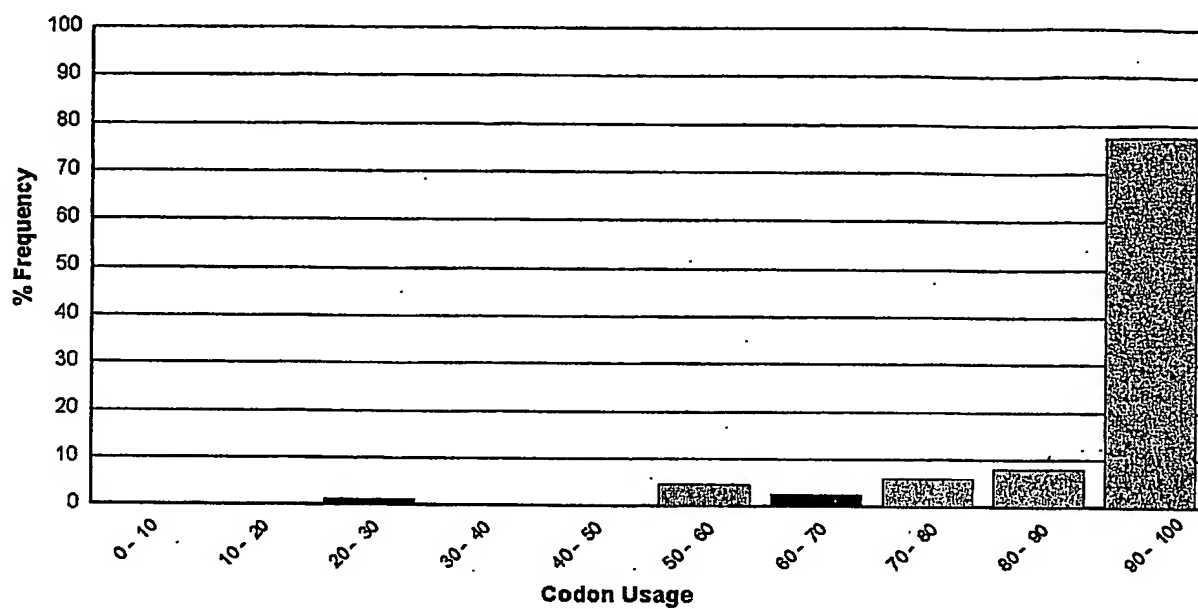


Fig. 7a

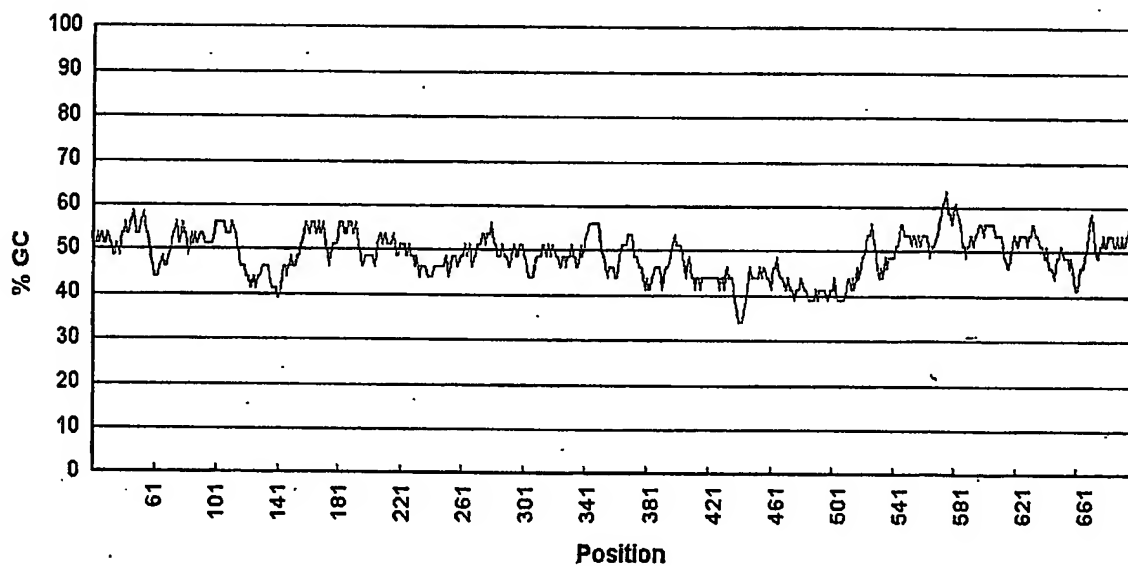


Fig. 7b

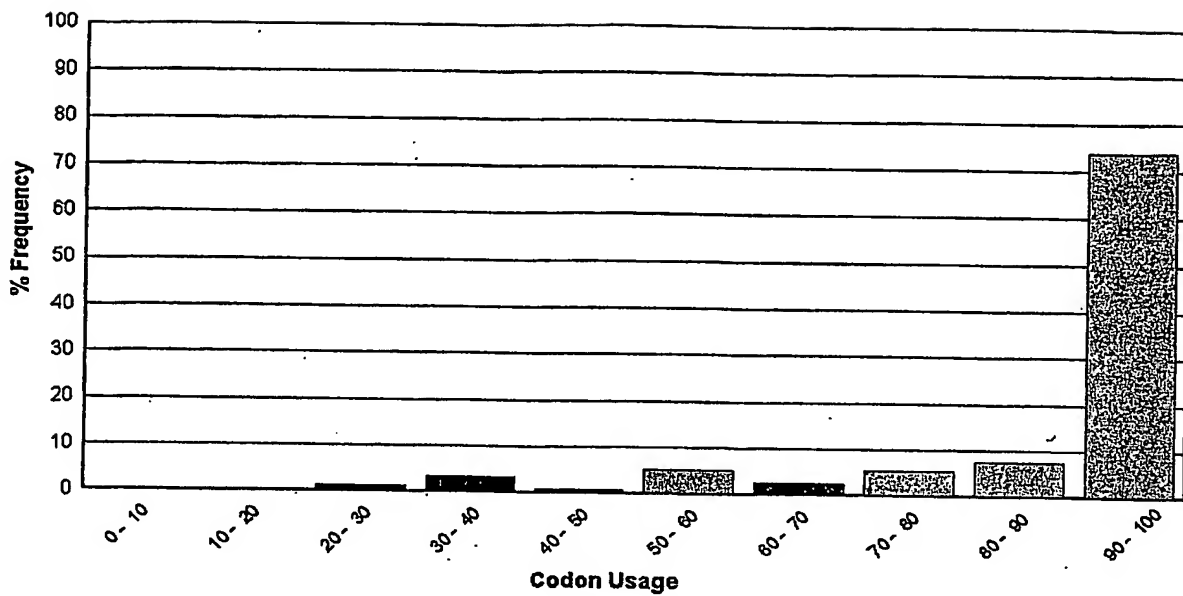


Fig. 8a

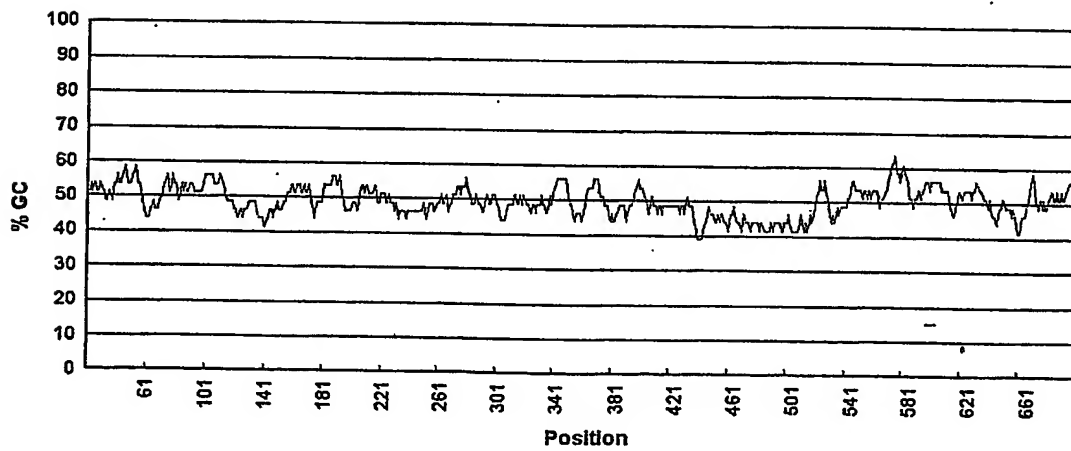


Fig. 8b